DOCUMENT RESUME

ED 197 962 SE 034 033

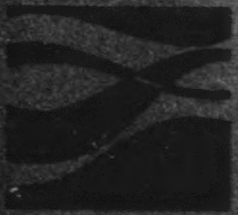
AUTHOR Harvey, John G., Ed.: Romberg, Thomas A., Ed.
TITLE Problem-Solving Studies in Mathematics. Monograph Series.
INSTITUTION Wisconsin Univ., Madison. Research and Development Center for Individualized Schooling.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 80
GRANT NIE-G-80-0117
NOTE 283p.: Not available in hard copy due to copyright restrictions. Contains light and broken type.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Behavior Theories: *Cognitive Processes: Discovery Learning: *Educational Research: Elementary Secondary Education: *Learning Theories: Literature Reviews: *Mathematics Education: *Problem Solving
IDENTIFIERS *Heuristics: *Mathematics Education Research

ABSTRACT

This monograph focuses on educational research on the processes and natures of problem-solving activities in mathematics. The first chapter presents an overview to both the field and the document itself. All of the studies reported reflect interrelated investigations carried out at the University of Madison-Wisconsin, as partial fulfillments of Ph.D. requirements in Mathematics or Curriculum and Instruction. Chapter two describes 31 studies carried out between 1969 and 1978, and divides the research into three categories: instruction in heuristics, assessment of problem-solving performance, and correlates and factors of problem-solving performance. The next four sections are reports of studies on teaching problem solving, chapters seven and eight detail investigations on assessing problem-solving performance, and the last three portions describe studies on establishing correlates and factors of problem-solving performance. (MP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****



PROBLEM-SOLVING STUDIES IN MATHEMATICS

Edited by John G. Harvey and Thomas A. Romberg

Wisconsin Research and Development Center for Individualized Schooling Monograph Series

University of Wisconsin-Madison School of Education

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

WRDC

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED197962

SE 034 033

© 1980 The Board of Regents of the University of Wisconsin System for the Wisconsin Research and Development Center for Individualized Schooling. This work was developed under a grant from the National Institute of Education, Department of Health, Education and Welfare. However, the content does not necessarily reflect the policy of that Agency and no official endorsement of these materials should be inferred.

Center Grant No. OB-NIE-G-80-0117

CONTENTS

Acknowledgments

Chapter 1
STUDIES ON MATHEMATICAL PROBLEM SOLVING: AN OVERVIEW
Thomas A. Romberg

Chapter 2
PROBLEM SOLVING IN MATHEMATICS: 1969-1978
John G. Harvey

Chapter 3
THE SMALL GROUP DISCOVERY METHOD: 1967-1977
Neil Davidson

Chapter 4
DEVELOPMENT OF A UNIT OF NUMBER THEORY FOR USE IN HIGH SCHOOL, BASED ON A HEURISTIC APPROACH
Shlomo Libeskind

Chapter 5
AN EXPLORATORY STUDY ON THE DIAGNOSTIC TEACHING OF HEURISTIC PROBLEM-SOLVING STRATEGIES IN CALCULUS
John F. Lucas

Chapter 6
A MULTIDIMENSIONAL EXPLORATORY INVESTIGATION OF SMALL GROUP-HEURISTIC AND EXPOSITORY LEARNING IN CALCULUS
Norman J. Loomer

Chapter 7
A STUDY OF PROBLEM-SOLVING PERFORMANCE MEASURES
Donald L. Zalewski

Chapter 8
DEVELOPMENT OF A TEST OF MATHEMATICAL PROBLEM SOLVING WHICH YIELDS A COMPREHENSION, APPLICATION, AND PROBLEM-SOLVING SCORE
Diana C. Wearne

Chapter 9
MATHEMATICAL PROBLEM-SOLVING PERFORMANCE AND INTELLECTUAL ABILITIES OF FOURTH-GRADE CHILDREN
Ruth Ann Meyer

Chapter 10
SEX, VISUAL SPATIAL ABILITIES, AND PROBLEM SOLVING
Ann Schonberger

Chapter 11
RELATIONSHIPS BETWEEN SELECTED NONCOGNITIVE FACTORS AND THE PROBLEM-SOLVING PERFORMANCE OF FOURTH-GRADE CHILDREN
Donald R. Whitaker

References

ACKNOWLEDGMENTS

The monograph editors would like to thank Robert Cavey, Teri Frailey, Marcia Grayson, Jean Padrutt, Mary Pulliam, Madeline Quigley, and Suanne Wamsley of Media Services at the Wisconsin Research and Development Center for Individualized Schooling for supplying the graphics and assisting in editing *Problem Solving Studies in Mathematics*. Our particular thanks go to Jean Padrutt who coordinated the project within Media Services and worked most closely with the editors; with her help the final editing of the monograph was easier and resulted in a volume which is much better than it would have been otherwise.

Chapter 1

Studies on Mathematical Problem Solving: An Overview

Thomas A. Romberg

"Problems worthy of attack, prove their worth by hitting back." (Hein, 1966)

Some mathematics students find joy in attacking worthy problems, and some mathematics teachers find joy in instructing their students on how to attack such problems. This monograph on problem solving addresses the following questions: "How can we teach problem solving know-how?"; "Who has problem-solving capabilities?"; and "What other intellectual abilities are related to that capacity?"

To introduce this monograph, I have chosen an example of a mathematics problem given to me to solve.

> Given intersecting spheres A and B with B passing through the center of A, find a formula for the surface area of B contained in A. (Polya, lecture notes, 1960)

I vividly remember when I perceived the solution to this problem. I was walking in the Quad at Stanford during the lunch hour after vainly struggling for at least a day to discover an appropriate relationship which might lead to a solution. In an instant, I realized that if the extreme cases of sphere B contained in sphere A and still intersecting it were considered, they had the same surface area. Although there was much work still to be done to prove my insight for a general case, I was convinced I had solved the problem. This incident, which occurred nearly 20 years ago, is only one of many I could relate which evolved from a series of problem-solving seminars offered by Professor George Polya of Stanford University for mathematics teachers sponsored by the National Science Foundation.

I chose this example for three reasons. First, while the roots of the individual studies reported here are a part of each author's background and training, all of contemporary mathematics education has been significantly influenced by George Polya and his writings on mathematical problem solving. In particular, *Mathematical Discovery* (Polya, 1962) was used as a reference book in courses taken or taught by all of the contributors to this volume. The above problem, assigned to me by Polya, is illustrative of the types of problems he used to teach problem solving. The strategy I used, i.e., looking at extreme cases, is one he advocates. His influence on me was considerable. Although I had a great deal of mathematics training, had worked as an applied mathematician, had taught high school and college mathematics, and had even solved a

few interesting mathematical problems, Polya changed my orientation when I took my first course from him in 1960. He clarified my thoughts about mathematics and the teaching of mathematics, and improved my problem-solving know-how. His books have done the same for us all.

Second, while for many educators mathematics consists of a large set of concepts and skills to be mastered, to most mathematicians the capability of solving problems that "hit back" is the essence of the discipline. As Polya (1962) stated:

> Solving a problem means finding a way out of a difficulty, a way around an obstacle, attaining an aim that was not immediately attainable. Solving problems is the specific achievement of intelligence, and intelligence is the specific gift of mankind: solving problems can be regarded as the most characteristically human activity. (p. vii)

While it is true that problem solving is an intellectual activity associated with all areas of inquiry, mathematics is one area where problems "worthy of attack" can readily be posed, and from such problems the intellect can practice problem solving. Thus, this monograph is limited to mathematical problem solving.

Third, this problem was assigned to a group of mathematics teachers, not mathematicians, psychologists, sociologists, or curriculum writers. For classroom teachers like myself who have experienced the exhilaration of solving a problem, a fascination grows in spite of the difficulty and frustration one often encounters in attempting to solve problems. Teachers become interested in how to teach the know-how (the strategies or heuristics) of problem solving to their students. Teachers would like students to enjoy the exhilaration that accompanies successful problem solving. Thus, one worthy educational problem is: "How does one teach problem-solving skills?". Furthermore, any teacher who has attempted to teach problem-solving strategies finds only a small group of students enjoying and being able to solve problems, while a number of students are totally frustrated. Teachers would like to identify those students who have an aptitude for solving problems. This involves both directly assessing problem-solving performance and identifying correlates of such performances.

Again, the emphasis reflected in this monograph parallels these two concerns for teachers: namely, the teaching of problem-solving heuristics and the identification of students with problem-solving aptitude.

The Chapters in This Monograph

It is important to see the nine studies reported in this monograph in relation to the extensive body of research literature on problem solving. In this introductory chapter I outline my approach to the study of mathematical prob-

lem solving and briefly discuss each study's location with respect to that outline. But, first, let me briefly summarize the other chapters of the monograph.

Chapter 2: Problem solving in mathematics, 1969-1978. In this chapter, prepared by John Harvey, 31 research studies in problem solving, conducted between 1969 and 1978, are described. These studies all fall into one of three categories: instruction in heuristics, assessment of problem-solving performance, and correlates and factors of problem-solving performance.

The next four chapters are reports of studies on teaching problem solving.

Chapter 3: The small group discovery method, 1967-1977. In this chapter Neil Davidson describes an instructional technique which he calls "the small group discovery method." After describing this method he details its initial tryout and the subsequent uses which have been made of it.

Chapter 4: Development of a unit of number theory for use in high school, based on a heuristic approach. Shlomo Libeskind discusses his development of a number theory unit based on a heuristic approach. This chapter presents the data which Libeskind gathered when he tried out the number theory unit with high school students enrolled in the Michigan State University Inner City Program.

Chapter 5: An exploratory study on the diagnostic teaching of heuristic problem-solving strategies in calculus. This landmark study by John Lucas is a pivotal chapter in the monograph. Many of the studies which are subsequently detailed depend upon the Lucas study and his description of the Polya problem-solving heuristics, the thinking aloud procedure, and the methodology for summarizing and analyzing the process-product data arising from use of that procedure. In addition Lucas's chapter describes his attempts to teach the Polya heuristics to college students in a calculus course.

Chapter 6: A multidimensional exploratory investigation of small group-heuristic and expository learning in calculus. Norman Loomer, using Lucas's refined procedures for gathering and analyzing process-product data, evaluates Davidson's small group discovery method for teaching Polya's problem-solving heuristics.

The next two chapters report studies on assessing problem-solving performance.

Chapter 7: A study of problem-solving performance measures. Donald Zalewski describes the development of a paper-and-pencil problem-solving instrument for seventh-grade students, the use of this instrument, and his attempts to correlate the results with data obtained using the thinking aloud procedure.

Chapter 8: Development of a test of mathematical problem solving which yields a comprehension, application, and problem-solving score. Diana Wearne traces the development of an instrument designed to measure the problem-solving performance of fourth-grade students. The chapter presents data regarding the validity and reliability of the resulting instrument and details the tryout of the instrument. This instrument was also used in the studies reported by Meyer and Whitaker.

The last three chapters are reports of studies on establishing correlates and factors of problem-solving performance.

Chapter 9: Mathematical problem-solving performance and intellectual abilities of fourth-grade children. This chapter, by Ruth Ann Meyer, reports an investigation of relationships between mathematical problem-solving performance and intellectual abilities. She gave 19 tests on intellectual ability and one on problem-solving performance, and used factor analytic techniques to isolate six factors related to problem-solving performance.

Chapter 10: Sex, visual spatial abilities, and problem solving. Ann Schonberger reports her investigation of sex differences, spatial ability, and problem-solving performance. In addition, this chapter reviews the research literature concerned with the relationships between spatial ability and sex differences.

Chapter 11: Relationships between selected noncognitive factors and the problem-solving performance of fourth-grade children. This chapter, by Donald Whitaker, details a study in which he investigated relationships between problem-solving performance of children and both children's and teacher's attitudes toward problem solving in mathematics.

An Approach to the Study of Problem Solving

In terms of approach, I have chosen to organize ideas about problem solving by using a basic stimulus-response framework (see Figure 1). One can discuss problem solving as task or stimulus specification (the observable characteristics of a worthy problem), as process (the distinctive cognitive processes used to attack a problem), or as product (the distinctive characteristics of the responses as a result of attacking a problem).

In all of the nine studies some attention was given to task specification. Problems in each study are assumed to be mathematical in nature and to require the use of mathematical concepts and skills to find a solution. Thus, this volume is not about the applications of mathematics to other problem situations. In particular, no study is about how to develop mathematical modeling skills. I recognize that mathematical modeling is an important ability. It undoubtedly has a close relationship with problem solving, but that is not the emphasis of this document.

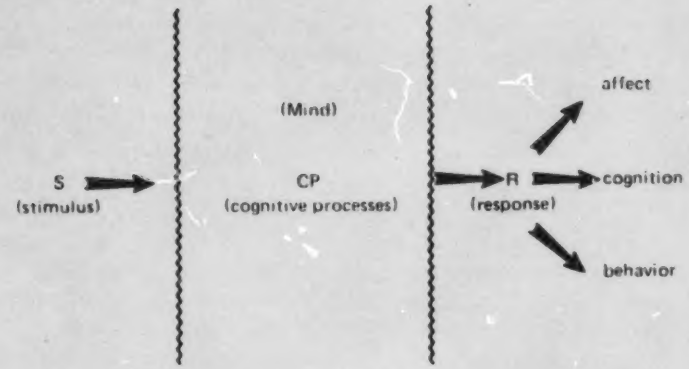


Figure 1. The Basic Stimulus-Response Framework.

In Schonberger's study (Chapter 10), the differentiation between spatial problems and quantitative problems is central to studying questions about how boys and girls approach problems in different ways. And for Wearne's study (Chapter 8), a hierarchical differentiation of questions about mathematical problems is of paramount importance.

Similarly, all of the studies make assumptions about the psychological processes used to attack problems. In the psychological literature on problem solving, two principal kinds of problem solving have been distinguished. The "trial-and-error" approach involves a series of successive approximations. The "insightful" approach involves a discovery of a meaningful means-end relationship underlying the problem (Ausubel, 1968). Only insightful problem solving is considered here. Insight may involve either a simple transposition of a previously learned principle to a new situation, or a cognitive restructuring and integration of experience to fit the demands of a designated problem. Characteristically, insightful solutions emerge suddenly. However, solutions are not always complete. They often appear after a protracted period of inauspicious search spent in pursuing unpromising leads.

Insightful problem solving is a type of meaningful discovery learning in which problem conditions are nonarbitrarily related to existing cognitive structure. Solving such problems involves going beyond the information given by transforming information, through analysis, synthesis, rearrangment, recombination, etc. The mathematical techniques we call heuristics, assumed to be useful in transforming information are those discussed by Polya (1945, 1954, 1962). In particular, see Lucas's analysis of Polya's heuristics (Chapter 5). What should be clear is that although psychological processes associated with problem solving are being examined, the studies reported here do not

attempt to clarify the intellectual processes that one uses when solving a problem. In essence, these studies are not basic psychological studies. However, in the last three chapters Meyer, Schonberger, and Whitaker examine the relationship of measures of other psychological factors to measures of problem-solving capability.

To assess the use of heuristics identified by Polya when solving problems, the coding procedures originally developed by Kilpatrick (1967) were followed. These coding procedures were for verbal protocols derived from students when instructed to "think aloud" while solving problems. To code his data Lucas (Chapter 5) adapted Kilpatrick's procedures for calculus students. Loomer (Chapter 6) and Zalewski (Chapter 7) then used variants of Lucas's coding in their studies. In particular, Zalewski used video recordings so that use of heuristics could be coded from visual as well as oral data.

Four papers in this monograph focus on teaching students to use heuristics for solving problems. Lucas and Libeskind rely on guided or arranged discovery. Davidson and Loomer, on the other hand, rely on small-group dynamics. In varying degrees all studies demonstrate that students can improve at solving problems. However, since all have given their subjects ample opportunity to solve problems, the long-debated "opportunity-to-learn" question in the literature is not clarified. Briefly, some psychologists, such as Ausubel (1963), have argued that because so few students are capable of solving problems, it is not a good use of time to try to teach all students problem-solving skills. This belief implies that those who are capable will develop those skills naturally. On the other side, Polya argues that problem solving, like other skills, needs to be practiced.

Finally, Whitaker, while not examining the teaching act itself, is interested in the attitudes teachers bring to the teaching of problem solving.

All of the studies consider *product* or *responses*. Each study examines whether problems are solved correctly or not, and if errors are made, the errors are classified. In particular, Zalewski and Wearne use the pattern of responses by individual students on instruments they developed to cluster the students. Zalewski's items were the basic set of items from which Schonberger selected items for her study. And, the instrument developed by Wearne was used by Meyer and Whitaker in their studies.

In summary, this monograph reports some interesting, interrelated studies conducted at the University of Wisconsin-Madison. All of the studies were carried out to partially fulfill the Ph.D. requirements in Mathematics or Curriculum and Instruction. Either Professors Harvey or Romberg chaired each thesis committee. All but one of the studies were partially supported by the Wisconsin Research and Development Center for Individualized Schooling.

Chapter 2

Problem Solving in Mathematics: 1969-1978

John G. Harvey

In 1969 Kilpatrick (1969, 1970) ably and comprehensively reviewed research in problem solving in mathematics. This chapter will update portions of that review for the years 1969 to early 1978.

Before beginning, it seems wise to briefly discuss the criteria used in selecting the studies described since this review will not be as comprehensive as Kilpatrick's. The problem-solving studies conducted at the University of Wisconsin from 1968 to 1977 and reported in this volume have implicitly or explicitly used the following definition of a mathematical problem and of mathematical problem solving.

> A *mathematical problem* is a situation which poses a question or defines an objective in light of some given information or conditions; the individual attempting to answer the question or meet the objective does not possess an immediate solution; hence the *solution process*, or act of solving a mathematical problem, requires active search, prior knowledge of mathematics, and a repertoire of heuristic strategies (Lucas, 1972, p. 10).

In addition the Wisconsin studies fall into three areas of problem-solving research: (a) instruction in heuristics, (b) measurement of problem-solving performance, and (c) correlates and factors of problem-solving ability. As a result this review only reports studies meeting these two criteria: The problems used were mathematical problems and the research is in one of the three areas named.

The second criterion is also used to organize the majority of this chapter. The next section will describe studies reporting attempts to teach heuristic strategies and the results of those attempts. The measurement of problem-solving performance will be the subject of section two. The third section details research which sought correlates and factors of problem-solving ability.

Instruction in Heuristics

Single Treatment Studies

Two chapters in this volume, those by Libeskind and Davidson, describe the initial tryouts of new instructional systems designed to teach problem-solving heuristics. Appropriately, neither Libeskind nor Davidson attempted to compare their new instructional system to "conventional" or to other innovative instructional systems; instead they focused their attention on

the components of their respective systems to determine if they functioned as planned. The resulting study can be termed a "single treatment study"; this part reports four other single treatment studies attempting to teach problem-solving heuristics.

In his study Gallo (1975) examined the role of two problem-solving processes. One of them, which he termed Integration, is the capacity to integrate other problem-solving processes into the sequence of operations required for problem solution; the second, termed Evaluation, is the capacity to judge whether an attempted solution is correct. These two processes and three other, unspecified problem-solving processes were taught to sixth-grade subjects. The treatment used was structured so that each of the processes could be taught in the context of computing the area of a triangle. This prevented inadvertently teaching the interrelation between the processes, and permitted the inclusion or omission of either or both of the processes of Integration and Evaluation. Gallo's results showed that when his subjects had learned both processes the solution rate was nearly perfect and when either was absent the solution rate did not exceed chance. The number of subjects, the length and duration of the treatment, the kind of problem-solving instruments employed, and the way the problem-solving instruments were used were not described in the abstract of this study.

An exploratory study by Dalton (1975) attempted to determine whether there were patterns in the thinking processes used by students of average or below average ability in mathematics. Next, it described the existing patterns, and determined the effects of "guiding questions" upon the thinking processes used and upon finding correct solutions. The subjects were 44 ninth-grade general mathematics students; they were assigned to an experimental and a control group of 22 students each. In both the experimental and the control group the students were asked to think aloud while solving three word problems; these individual thinking aloud interviews were tape recorded. In the experimental group the students were asked "guiding questions" during the interview. In his abstract Dalton did not report the length of the thinking aloud interviews, the number of "guiding questions" asked of subjects in the experimental group, or the way the data were analyzed. He did report that the tape recordings were transcribed and coded, that the errors were analyzed, and that his general observations of the subjects were used. He concluded that there were patterns in the thinking processes of his subjects; two modes of thinking, deduction and trial-and-error, were used; and subjects who used trial-and-error tended to be more effective problem solvers. Dalton further states that "the effects of asking students 'guiding questions' were not determined conclusively."

The study by Kantowski (1977) is similar to Davidson's in that it spanned 8 months of a school year and to those of Libeskind, Lucas, and Loomer in that the treatment embodied the heuristics identified by Polya

(1957, 1962, 1965). The subjects in this four-phase, clinical investigation were eight high-ability ninth-grade algebra students (four females, four males). The first phase was an eight problem pretest. The second phase was readiness instruction (three lessons per week for 4 weeks) intended to acquaint the subjects with heuristic instruction and to introduce them to using heuristics in problem solving. This phase concluded with a test. The third phase, 4 months in duration, was heuristic instruction in geometry. There were three units of geometry content. Each unit consisted of six initial instructional episodes, a midunit test, six more instructional episodes, and an end-of-unit test. The fourth phase was a two-part posttest. One part consisted of geometry and verbal problems; the other part, of prerequisite knowledge needed to solve the geometry and verbal problems. The number of items in the phase-two test, the phase-three tests, and the two parts of the posttest were not given. All of the tests were individually administered. During each test the subjects were encouraged to think aloud as they solved problems. Each problem-solving interview was tape recorded, the subjects' protocols were analyzed from these tapes using a modification of the coding scheme developed by Kilpatrick (1968), and a process-product score was assigned to each problem solution.[1]

Using the process-product score Kantowski (1977) calculated a median decimal score for each of the eight subjects. Then, for each subject, she determined the percentages of problems in which the problem-solving processes were used. Percentages were calculated for problems with scores above and with scores below the median. Based upon these data Kantowski reported the following: (a) 59 to 95% of the problem solutions with scores above the median showed evidence of the use of goal-oriented heuristics, while at most 52% of the problem solutions with scores below the median showed indication of their use; (b) the tendency to use goal-oriented heuristics increased as problem-solving ability developed; (c) the percentage of problem solutions indicating the use of goal-oriented heuristics ranged between 14 and 72% with a median of 36% on the pretest, and between 14 and 100% with a median of 72% on the posttest[2]; (d) successful problem solvers manifested regular patterns in using the processes of analysis and synthesis, and there is an interrelationship between these regular patterns and using goal-oriented heuristics; and (e) the subjects seldom used the heuristic of looking back.

In his study Vos (1978) chose the following three key organizers: drawing a diagram, approximating and verifying, and constructing a chart. He hyphothesized that these organizers would potentially increase success in

---

[1]The Kilpatrick coding scheme and the way in which process-product scores are assigned are more fully described in Chapter 5.

[2]The range does not seem to be a good representation of change from pretest to posttest in this case. If the posttest score of one subject is deleted, then the range is from 72 to 100% with a median of 72% ($N = 7$).

problem solving. For each organizer he developed an instructional treatment of six presentations. Using a pretest-posttest design Vos taught the three treatments to 21 randomly selected subjects from grades six, seven, and eight (seven at each grade level) over a 14-week period. The pretest instruments consisted of a test of mathematics ability, a learning style inventory (a modification of Learning Style Inventory-A [Kolb, Rubin, & McIntyre, 1974]), and a problem-solving test. The posttest instruments were a practical judgment test (developed from Tate & Stanier, 1964), a Problem Solving Decision Test (Vos, 1976), and a problem-solving test. The problem-solving pretest and posttest were individually administered and consisted of three and six items, respectively. Each subject was instructed to think aloud during the tape-recorded interviews. Using the process coding scheme developed by Kantowski (1977), a process-product score was assigned to each solution for the problems in the problem-solving tests. Based on the pretest and posttest data Vos concluded the following: (a) each of his instructional treatments was successful, (b) his subjects did use the three key organizers in problem-solving situations, (c) there was a relationship between effective application of the key organizers and success in problem solving, and (d) for eighth-grade subjects, there was a strong relationship between problem-solving success and practical judgment.

Treatment Comparison Studies

At present the more conventional educational research paradigm is to compare the effects of one treatment to the effects of another. The studies reported here are of that kind. However, the studies are further subdivided into those in which a heuristic treatment was compared to a conventional one, and those in which more than one heuristic treatment was used.

Heuristic vs. conventional instruction. Leggette (1974) attempted to determine if instruction in heuristic processes would increase the problem-solving performance of capable, but poorly prepared college freshmen. Four intact classes, totaling 70 college freshmen, were assigned to an experimental group and a control group. Two instructors were randomly assigned to one of the experimental and one of the control classes. Both the control and experimental treatments lasted 9 weeks; during the treatment period the control classes "followed normally scheduled class procedures." During the first week of the treatment period each experimental class received 3 hours of instruction on problem-solving processes; for the rest of the treatment period those classes were taught mathematics using a problem-solving approach. The Basic College Mathematics Problem-solving Test and the Aiken Revised Mathematics Attitude Scale (Aiken, 1963) were administered to both groups as pre- and posttests. There were no significant differences ($p = .01$) in problem-solving performance between the experimental and the control group. Analysis of variance was used to determine if there were significant differences between the problem-solving mean gain scores and the attitude mean gain scores of the two treatment groups. It was concluded that: (a) the experimental treatment increased the problem-solving ability of capable, but poorly prepared college

freshmen more than the control treatment; (b) the experimental treatment "should cause students to develop a better attitude toward mathematics"; (c) these freshmen could be taught the structure of problem solving without affecting the amount of mathematics content taught; and (d) an undergraduate mathematics course with a unit on problem solving could be introduced.

Post and Brennan (1976) compared a general heuristic treatment with normal instruction in tenth-grade geometry. In the spring of 1972, 94 tenth-grade students were pretested using an investigator-developed problem-solving instrument. The subjects' scores were rank ordered, and pairs of persons with adjacent or coincident pretest scores were formed. One student from each of the resulting pairs was randomly assigned to the experimental group and the other to the control group. A median split divided both groups into high and low cells. The experimental classes were given teacher-directed large-group instruction which emphasized solving problems using the experimenters' General Heuristic Problem-solving Procedure. The control group continued normal instruction in geometry. Post and Brennan did not specify length of treatment. The same instrument was administered for both the pretest and the posttest. Two-way analysis of variance was used to compare the posttest means of the experimental and control groups. There were no significant treatment effects or interactions. There was a significant difference ($p < .01$) due to ability level.

Lee (1977) sought to improve the heuristic problem-solving behaviors of fourth-grade students in his exploratory study. Using teachers' recommendations and students' performance on two Piagetian problems (Equilibrium in the Balance and Oscillation of a Pendulum), 16 subjects were selected for this experiment: eight average achievers who met Piaget's criteria of II-A cognitive level on both problems and eight high achievers who met Piaget's criteria of II-B cognitive level. Two groups of equal size, an experimental and a control group, were formed by random assignment of subjects within a stratum. The experimental group was instructed on the use of heuristics when solving word problems. Although the treatment given to the control group was not specified in the abstract, it seems reasonable to asume that they continued to receive their usual instruction in fourth-grade mathematics. The experimental treatment lasted for 8 weeks; during that time there were 20 instructional sessions of 45 minutes each. Pre- and posttests were given to both treatment groups. The pretest consisted of two problems; the posttest, six problems. Four weeks after the end of the treatment period the experimental group was given two additional problems to solve. Tape recordings were made during the individually administered testing sessions. In addition subjects' worksheets and the investigator's remarks were collected. On the posttest subjects in the experimental group solved 35 of the 48 problems (73%); control group subjects solved 3 of the 48 problems (6%). Subjects in the experimental group solved 80% of the problems presented to them during the 4-week follow-up testing. The investigator reported the following: (a) there was no change in

the use of heuristics by control group subjects while there was a noticeable increase in their usage by experimental group subjects, (b) subjects in the experimental group "were able to select an appropriate heuristic for nearly all the post-experimental interview problems," and (c) there was a difference in the use of heuristics between those subjects classified as meeting Piaget's criteria of II-A cognitive level and those meeting II-B.

Ledbetter (1978) attempted to isolate an aptitude-treatment interaction in her study of heuristic problem solving. A total of 84 college freshmen were randomly divided into an experimental and a control group. During the 10-week treatments experimental subjects received instruction on problem solving and the use of heuristic strategies, while the control subjects were instructed in college algebra and trigonometry. All subjects took five ability pretests. A Solomon four-group design provided data on problem-solving performance and problem-sorting schemes (Silver, 1978) for approximately half of the experimental and control group subjects. There were posttest measures of problem-solving performance, algebra and trigonometry performance, and problem-sorting schemes; the posttest instruments were administered to all subjects. The nine-item problem-solving test included problems solved by three heuristic strategies (algebraic symbolism, contradiction, and pattern generation) and incorporated three contextual cues (triangle, number, and word problems). The problem-sorting schemes data were gathered using a problem-similarity questionnaire that required subjects to rate each of nine pairs of problems on a continuous similarity scale. Experimental subjects outperformed control subjects on the problem-solving posttest ($p < .01$), while the contrary was true on the algebra-trigonometry posttest ($p < .001$). A complete-link clustering analysis of the problem-sorting scheme data indicated that few differences in dominant clustering schemes could be observed. A heuristic sorting score was significantly correlated with problem-solving performance ($p < .04$); the correlation coefficient was not stated. A hierarchical clustering analysis of the ability test data isolated four homogeneous ability profile groups. Analysis of variance showed that the heuristic sorting score was related to ability profile group ($p < .01$) and to treatment group ($p < .001$). Subjects in the experimental group received higher sorting scores. To test for aptitude-treatment interactions, the problem-solving posttest was divided into three subtests corresponding to the three heuristic strategies taught to the experimental group. Results showed that only one of the ability profile groups performed significantly better (the p-level was unspecified) across all three subtests following treatment.

Like the four studies just described, the next two studies attempted to determine the effects of teaching problem-solving heuristics to their subjects. However, the two remaining studies also attempted to determine the effects of heuristic instruction on students of the subjects.

The following three outcomes were investigated by Lipson (1972): the subjects' problem-solving performance, the problem-solving performance of children taught by the subjects, and the subjects' teaching behavior. The subjects for this study, 43 senior mathematics majors enrolled in a secondary school mathematics methods course, were divided into three cells: students who had participated in an experiment as freshmen and had received instruction on heuristics, students who had participated as freshmen and had not received heuristics instruction, and students who had not participated as freshmen. The majority of the subjects were in the first cell. Half of the 43 subjects were assigned to the experimental treatment, a seminar on heuristics, and the other half to the control treatment, continued participation in the regular methods class. The abstract did not describe the way subjects were assigned to treatments or cells. It did explain that "the subjects were partitioned into six subsamples on the basis of treatment and freshmen experience." Problem-solving pre- and posttests were administered to all subjects; there was an intervening treatment period whose length was not specified. While the subjects were student teachers, they administered problem-solving pre- and posttests to their students. During the same period, trained observers recorded the heuristic activities of the subjects as student teachers. Analysis of variance of the pretest-posttest gain scores demonstrated that there were no significant differences between the six subsamples. When the 43 subjects were divided into groups who scored low, medium, or high on the pretest, a two-way analysis of variance yielded a significant difference ($p < .01$) favoring the subjects in the experimental treatment. A one-way analysis of variance was used to compare the pre- and posttest means of classes taught by student teachers; there were no significant differences ($p < .01$) between the classes of the student teachers from the six subsamples. Scheffe's method of multiple comparisons located several significant contrasts ($p < .01$). These contrasts were not described in the abstract, but it was concluded that, on the average, classes taught by student teachers who participated in the experimental treatment had gained more in problem-solving performance. There were too few instances of observed heuristic teaching to permit statistical analysis. It was stated that the subjects who participated in the experimental treatment and who had had instruction in heuristics as freshmen showed more instances of heuristic teaching. A higher pretest score was related to greater heuristic behavior as a student teacher.

A similar study has been conducted by Tubb (1975). In his study mathematics graduate teaching assistants were trained in heuristic questioning strategies, Flanders' Interaction Analysis, or both. Problem-solving performance of the graduate students and their calculus students was measured. Several positive results are stated. However, the problems used in this study do not satisfy the definition of a mathematical problem, and thus, the study does not meet the criteria established for this review. Therefore, the study will not be further described.

Comparison of heuristic treatments. The studies in this section compare one or more heuristic treatments. They are distinguished from the previously described studies because the treatments in this group are usually better specified.

Pennington (1970) examined the following two approaches to the teaching of heuristic strategies: the Behavioral Strategy Treatment and the Conceptual Strategy Treatment. Strategies in the behavioral treatment were specific, logical steps for problem solution, while conceptual strategies were organizing principles. In addition two types of problem-solving practice were considered, Selection and Reception. At the Selection level subjects arranged their own learning sequence during training; at the Reception level there was a predetermined learning sequence. Four instructional treatments were derived by pairing each heuristic approach with each level of practice. The content of each treatment consisted of structure problems developed "according to a system specified by mathematical group theory." The subjects for this study were sixth-grade students who, prior to participating in the experiment, were taught modular arithmetic addition. In addition to the four instructional treatment groups, subjects were also assigned to a control group. The number of subjects in each group, the way subjects were assigned to groups, and the length of the instructional period were not given in the abstract. Problem-solving performance was measured by three acquisition tests administered during training, and by learning and transfer tests administered afterward. The content of the tests was not specified. The difference between the treatment and control groups reached the (unspecified) predicted level of significance on two of the three acquisition measures, and on all of the learning and transfer measures. There were no other significant differences.

Foster (1973) hypothesized that a student who successfully programmed a computer to solve a series of mathematical problems would develop his or her problem-solving ability. For this posttest-only experiment, Foster defined four treatments by specifying the kind of supplementary aids used in each. They were no aids (Treatment 1), flow charts only (Treatment 2), computer only (Treatment 3), and computer and flow charts (Treatment 4). The subjects for this experiment were three intact eighth-grade classes of 24 students each. After dividing each class into two equal strata using reading ability, stratified random sampling was used to assign subjects to one of the four 12-week treatments. The posttest was a 48-item, experimenter-constructed test of nine problem-solving behaviors. A two-way analysis of variance of mean performance on the posttest was used to determine the effects due to treatment, class, and reading level. Significant F-values results for reading within treatment ($p = .01$) and treatment × class ($p = .05$). Pair-wise comparisons, using Scheffe's t-statistic, failed to show if these significant F-values were within one treatment or a reading cell within a treatment. An analysis using Dunnett's t-statistic revealed that the mean performance of those using the computer only was significantly greater ($p = .05$) than those using neither

the computer nor flow charts. The mean performance of the four treatment (T) groups had the directional order: T1 < T2 < T4 < T3.

In a study similar to Pennington's (1970), Smith (1973) compared the effects of giving general versus specific heuristic advice. The general heuristic taught was the planning heuristic successfully used by General Problem Solver (Ernst & Newell, 1969); the specific heuristic taught applied best to the task being studied. The investigator developed three programmed study booklets on finite geometry, Boolean algebra, and symbolic logic, three tests covering the booklet material, and two transfer tests. The subjects, 176 college students who had taken 2 years of high school mathematics, were assigned to two treatment groups, the general heuristic treatment and the specific heuristic treatment. Within those treatments nine additional factors were identified (three orders of booklet presentation × three orders of booklet test administration). Each treatment lasted for 3 weeks; the five investigator-designed tests were administered during the fourth week of the study. Information concerning the subjects' problem-solving methods was gathered by means of interviews and questionnaires after the testing period. The interview procedures and the content of the questionnaires were not described in the study abstract. The data were analyzed using a three-way analysis of variance. Subjects in the specific heuristic treatment group solved significantly more ($p < .001$) logic problems and completed the Boolean algebra and logic tests significantly faster ($p < .05$ and $p < .01$, respectively) than did subjects in the general heuristic treatment group. There were no significant differences between the treatment groups on the number of transfer problems solved and the time required to solve them. There were no main effects for order and no interactions. The questionnaire and interview results showed that one- to two-thirds of the subjects used the heuristics taught to them when completing a given learning test and that very few used the heuristics taught to them on the transfer tests.

Training in heuristics was approached differently by Goldberg (1975). She studied the effects of training in heuristics on the ability to write proofs in number theory. Goldberg developed two sets of programmed materials; one set provided heuristic instruction and the other did not. Three treatments were designed: Treatment XT used the heuristic programmed materials and classroom instruction reinforced those heuristics; Treatment X used the heuristics programmed materials and classroom instruction did not provide reinforcement; Treatment C used the nonheuristic programmed materials and classroom instruction did not teach heuristics. Nine intact classes were randomly assigned to these treatments; each class met for 75 minutes, twice weekly for 6 weeks. During seven of these class meetings subjects worked on the programmed materials; during the remaining five classes appropriate classroom instruction was provided. At the end of the treatment period the following four posttests were administered: a 25-item test of basic concepts studied (Concepts I), a test requiring the construction of proofs (Proofs I), a

questionnaire designed to determine attitude toward the programmed materials, and the Childhood Attitude Inventory for Problem Solving (Covington, 1966). Five weeks later two tests, Concepts II and Proofs II, were administered. Parts of these two tests were parallel to Concepts I and Proofs I; the remaining parts tested mastery of material studied subsequent to the treatment period. There were no significant differences between treatment groups in subjects' understanding of the basic concepts or their ability to construct proofs. Students responded that the nonheuristics programmed materials were more helpful, easier, and generally more appealing than the heuristic materials. The results of the attitude inventory showed that students given Treatment C had a more positive attitude toward the nature of the problem-solving process than subjects who received Treatments X or XT. Subjects given Treatment XT had more positive attitudes than subjects who received Treatment X.

Pereira-Mendoza (1976a, 1976b) taught students to apply at least one of two heuristics, examination of cases and analogy, to mathematical problems. He also investigated the differences between learning these heuristics alone or learning them in concert with mathematical content. He specified three levels of treatment (heuristics only [H], heuristics and content [HC], content only [C]) and three instructional vehicles (algebraic, geometric, mathematically neutral). Nine self-instructional booklets were designed which corresponded to each of the treatment by vehicle combinations. The subjects (N = 294) were tenth-grade boys in an all-male Canadian high school; they were randomly assigned to one of the nine groups. At the end of the 10-day instructional period two transfer tests, one algebraic and one geometric, were administered to all subjects. After eliminating tests on which judges could not reach scoring agreement and then equalizing the group sizes by random elimination of test scores, data from 189 subjects (21 per group) were analyzed using analysis of variance. On the algebraic test the H treatment groups, and on the geometric test the HC and H treatment groups scored significantly higher than did the C treatment groups. The probability level was not specified. There were no significant differences between the HC and C treatment groups on the algebraic test or between the H and HC treatment groups on the geometric test. An analysis of the pattern of heuristic application revealed that both heuristics were employed on the algebraic test and there was little evidence of the use of analogy on the geometric test. There were no significant differences between the instructional vehicles.

Vos (1976) compared three instructional strategies for promoting the use of five problem-solving heuristics. The heuristics were (a) drawing a diagram, (b) approximating and verifying, (c) constructing an algebraic equation, (d) classifying data, and (e) constructing a chart. In the reception treatment the subject was given only the problem task. In the list treatment the subject was given the problem task and, after some time had lapsed, was required to read a checklist of desirable problem-solving behaviors. Next, the subject was instructed in specific problem-solving behaviors that could help

solve the given problem task before returning to it. In the behavior treatment subjects were first instructed in specific problem-solving behaviors which would help solve the subsequently given problem. The 33 subjects were students in six ninth-, tenth-, and eleventh-grade mathematics classes at a private high school. The mathematics classes and the numbers of subjects in those classes were Algebra II (25), Geometry (29), Math Survey (21), Elementary Algebra (8), and Algebra I (50). Using a one-factor, randomized complete block design with mathematics class as the blocking variable, subjects within classes were randomly assigned to one of the three experimental treatments. For each experimental treatment, instruction consisted of investigator-developed, self-instructional materials supplemented by a teacher. Over a 15-week period 20 problem tasks of 20 minutes each were given. Pretest data consisted of scores from the Sequential Tests of Educational Progress (STEP) (Cooperative Test Division, 1972) forms 2A and 3A, Mathematics Part II. Posttest instruments were STEP, forms 2A and 3A, Mathematics Part I and an investigator-constructed Problem Solving Approach Test (PSAT) and Problem Solving Test (PST). At the .05-level there were no significant differences between treatments except for subjects enrolled in Math Survey on Part I of PSAT, a direct measure of the five problem-solving heuristics taught.

Gifted high school students were the subjects for the study conducted by Hall (1976). He designed and validated a checklist of heuristics involved in formulating problems from situations, and used this checklist to rate performance of gifted students on situational problems before and after instruction in situational problem solving. In addition to the situational heuristic checklist a planning heuristics checklist, compiled from Polya's list (1957), was also developed. A total of 156 superior secondary school subjects, comprising 39 four-person teams, was randomly assigned to one of three treatment groups: situational heuristics, planning heuristics, and control. The length of the treatments, the nature of the posttest administered, and the data analysis procedures were not described in the study abstract. The results were that on situational problems the subjects in the situational heuristics treatment group gave significantly more heuristics than the control group ($p < .001$) and the planning heuristics treatment group ($p < .001$). On these same problems the planning heuristics treatment group gave significantly more ($p < .05$) heuristics than the control group. On "well-defined problems" the planning heuristics treatment group gave significantly more ($p < .05$) heuristics than the control group.

McClintock (1978) compared three treatments in his study: G_1 which taught calculator usage, Algebra I content, and problem solving; G_2 which taught problem solving and Algebra I; and G_3 which taught calculator usage and Algebra I. The subjects were average ability Algebra I students from a private girls' school. The majority were from middle to upper middle class, Anglo or Cuban families. The subjects had been randomly assigned to one of three classes; there were 10, 17, and 9 students in treatment groups G_1, G_2,

and G_3, respectively. The treatments lasted from mid-March until the first week of June. Pretests and posttests were administered to gather data on algebra achievement, inductive reasoning, and deductive reasoning. The tests used were the Lankton First-Year Algebra Test (Lankton, 1965), the Necessary Arithmetic Operations Test and Nonsense Syllogisms Test from the ETS Kit of Reference Tests for Cognitive Factors (French, Ekstrom, & Price, 1969a, 1969b), and an investigator-developed number sequence test. These data were analyzed through analysis of covariance, with the pretest data being used as the covariate. In addition pre- and posttest problem-solving data were collected from eight subjects from each of the treatment groups. These subjects were in the upper half of their treatment groups on the other pretest measures. Six pretest and eight posttest problems were given to each subject; these problems were solved while the subjects thought aloud. The pretest and posttest interview sessions were tape recorded and the taped protocols were analyzed using a process coding scheme similar to the one used by Kantowski (1977). The analysis of covariance revealed significant differences between treatment groups on the Lankton First-Year Algebra Test ($p < .05$) and the Nonsense Syllogisms Test ($p < .01$). There were no significant differences between treatment groups on Necessary Arithmetic Operations or the number sequence test. The adjusted mean performance of the three treatment groups had the directional order $G_1 > G_3 > G_2$ and $G_2 > G_1 > G_3$ on the Lankton First-Year Algebra Test and the Nonsense Syllogisms Test, respectively. Analysis of the protocols indicated that all of the subjects found the pre- and posttest problems were difficult to solve, there was a relationship between heuristic processes and productive inferences, in approximately 83% of the problem-solving sessions subjects employed systematic trial-and-error, and there was a marked increase in the use of algebraic equations between pretest and posttest.

Assessment of Problem-solving Performance

Instruction on problem-solving heuristics and assessment of problem-solving performance are equally important to research on problem solving. The studies described in this section assess and describe problem-solving performance. The section will be divided into two parts; those using thinking aloud to assess performance and those using techniques other than thinking aloud.

Thinking Aloud Assessment

Fuller (1972) sought to determine if students use different methods of solving mathematics problems when under and not under time constraints. Sixty-four subjects of average and above average intelligence were individually administered two problem-solving tests. On one test subjects had 3 minutes to solve each problem, and on the other test they were told they could have as much time as they needed. Subjects thought aloud during the tape-recorded problem-solving interviews. The recorded protocols were analyzed by the in-

vestigater; in the coding system she used, problem-solving processes were grouped into four categories: reading the problem, rereading the problem, deduction, and trial-and-error. To determine if a subject changed problem-solving methods between the two tests a pattern of problem solving was computed for that subject on each test. The two patterns were compared by a contingency table and the X^2-statistic. No significant differences were found; no trends in the changes between the two patterns could be identified.

Schwieger (1974) identified a theoretical model for analyzing mathematical problem solving which consisted of eight basic abilities. Face validity for the model was obtained by generating operational definitions of each ability and by gathering the comments and opinions of mathematics educators and mathematicians after giving them a list of the abilities together with their definitions, descriptions, and examples. Finally, a collection of mathematical problems from the areas of arithmetic, algebra, and geometry were used in problem-solving interview sessions with secondary school, undergraduate, and graduate students. The total number of students interviewed, the number of students at each level, the number of problems given to each student, and the length of the problem-solving interview were not described in the abstract. During the problem-solving interviews students were asked to think aloud. The resulting protocols were tape recorded and analyzed. The analysis indicated that the basic abilities of the model were "necessary and sufficient for explaining the observed problem solving process."

In his study Webb (1975a, 1975b) explored the use of problem-solving processes by high school students. The subjects were forty second-year algebra students (20 males, 20 females). They were asked to think aloud while solving eight problems from the experimenter-developed Problem Solving Inventory (PSI) of mathematical problems including the areas of geometry, algebra, and analytic geometry (Kulm, 1977). Sixteen pretest measures of cognitive and affective variables were administered. These variables included mathematics achievement, attitudes toward mathematics, spatial ability, verbal ability, reasoning ability, and problem-solving ability. A coding scheme adapted from Kilpatrick (1968) was used to record the tape-recorded protocols from the thinking aloud interviews; this scheme yielded a total score on the PSI and the frequency with which each of the problem-solving processes was used. Principal component analyses were performed separately on the pretest and process scores. A regression using the component scores as the independent variables and the total PSI score as the dependent variable showed that the Mathematical Achievement component accounted for 50% of the variance in the total scores. Heuristic strategy components, a subset of the process components, accounted for an additional 15% of the variance. Of the 10 heuristic strategies tested, eight were used to solve one or two problems. No sex differences were found. Overall it was concluded that better problem solvers use a wider range of strategies and techniques than do poorer problem solvers.

In a study similar to Webb's (1975a, 1975b), Gimmestad (1977) explored the processes used by community college students. Subjects ($N = 60$) were randomly selected from mathematics students attending two community colleges in Colorado. During each 1½ hour interview measures of IQ, mathematics achievement, conceptual tempo, and mathematical problem solving were administered. Subjects thought aloud while solving the eight problems on the investigator-developed mathematical problem-solving inventory (Kulm, 1977). The interviews were tape recorded and process coded. Gimmestad reported that the most popular processes with the subjects were deduction, trial-and-error, and equations. Significant correlations ($p = .05$) were found between total problem-solving score and use of the processes of exploratory manipulations ($r = -.34$), successive approximation ($r = .37$), and deduction ($r = .30$). Conceptual tempo, age, sex, and IQ were not significantly related to mathematical problem solving performance, but a significant correlation ($p = .05$) was found between mathematics achievement and mathematical problem-solving performance.

Blake (1977) attempted to determine the effects of problem context and the degree of field independence upon processes used in solving mathematical problems. Subjects were 40 eleventh-grade Algebra II students randomly selected from students in 14 classes; they were of average ability for students enrolled in their program (IQ range: 115-125). Subjects were matched using their scores on Witkin's Embedded Figures Test (Witkin, 1950). One subject in each pair was randomly assigned to one of the testing groups. One testing group was given five mathematical problems in a real world setting; the other group was given the same problems in a mathematical setting. Subjects were instructed to think aloud as they solved these problems during individual tape-recorded interviews. The protocols were coded using a system based on a mathematical problem-solving model by MacPherson. Blake found that problem context is unrelated to heuristics and the degree of field independence had a marked effect upon the use of heuristics and the number of correct solutions. Field independent subjects demonstrated use of a greater variety of heuristics ($r = .33$), more willingness to change their mode of attack ($r = .27$), and a greater number of correct solutions ($r = .30$). Both the total number and the number of different heuristics used accounted for a significant amount ($p < .01$) of the variance in the number of correct solutions. In particular, the use of heuristics accounted for an additional 21% of the variance not accounted for by core procedures (algorithms, diagramming, equations, and guessing). Changing mode of attack was significantly related ($p < .01$) to obtaining a correct solution.

The following seven cognitive processes were studied by Hollowell (1977): (a) understand the problem, (b) recall from memory, (c) formulate a hypothesis or general idea for problem solution, (d) attempt to find a provisional solution or develop a method of solution, (e) check against solution model or general form of answer, (f) verify provisional solution correct, and

(g) reject provisional decisions. Subjects were 30 high school juniors who thought aloud while solving three mathematical problems. One of these problems did not require specific algebraic or geometric knowledge, one was an algebra word problem, and one was a geometry proof. The investigator found that the problem-solving sequences for the three problems, while similar, had some important differences. The recall process (b) appeared more frequently in the process coding sequence for the geometry problem than in the sequences for the other two problems. In the sequences for the algebra word problem a rejection (g) tended to be followed by a new attempt at provisional solution (d). For the other two problems a rejection tended to be followed by the formulation of a new hypothesis (c). Total number and kind of processes used did not appear to be related to success or failure.

Ortiz-Franco (1978) hypothesized that: (a) the relationship between mathematical problem solving and reading ability, mathematics achievement, and reasoning was different for Chicano students and Anglo students; (b) reading, mathematics achievement, and performance on his problem-solving inventory (PSI) are significantly related to field dependence; and (c) the use of trial-and-error processes differentiates better than field dependence between problem solvers. The subjects were 40 Chicano students who had not taken an algebra class. Half of the subjects (9 males, 11 females; mean age 14.93 years) were dominantly Spanish speaking, and half (10 males, 10 females; mean age 14.38 years) were dominantly English speaking. Pretests of mathematical achievement, reading achievement, reasoning, field dependence, divergent thinking, and anagrams were given to each subject; these tests were in the subject's dominant language. The problem-solving inventory was administered in individual interviews where the subjects were instructed to think aloud. The investigator reported a significant correlation ($p < .01$) between problem-solving performance and mathematical achievement; the correlation coefficient was not reported in the abstract. There were no other significant differences.

Problem-solving Assessment Not Dependent on Thinking Aloud

Many of the studies in the heuristic instruction section and all of the above studies have depended on the thinking aloud procedure to assess problem-solving performance. This procedure is easily the most popular one at the present time. However, there are some disadvantages and some serious, unanswered criticisms of the procedure. The most serious disadvantage is the amount of time it takes to interview each subject individually and to have an examiner present at those interviews. A second disadvantage is the difficulty of training persons to reliably code the resulting audio- or videotapes of subjects' behavior. As several studies in this volume demonstrate, persons can be trained to reliably code the process behaviors. Therefore, reliable coding is a disadvantage and not a criticism of the procedure.

Criticisms include the following: (a) subjects may not report all of their thoughts, but only those which are "safe" or "acceptable"; (b) the problem-solving processes used during thinking aloud interviews may be different from those the same subject would use to solve the same problem while not verbalizing; (c) the equipment required to record a thinking aloud interview may distract the subjects; (d) it is very difficult, if not impossible, to employ the procedure successfully when the subjects are young; and (e) at present, the resulting process code data must be radically altered to analyze data from thinking aloud interviews (see Flaherty, 1973; Hallgren, 1976). Thus, in the years from 1969 to 1978, four studies, including those by Zalewski and Wearne reported in Chapters 7 and 8, have attempted to find other ways to assess problem-solving performance.

One important aspect of mathematical problem solving, the construction of valid proofs, was investigated by Lester (1973, 1974). A group of 19 public school children was randomly chosen from grades 1-3 (Group A_1), 4-6 (Group A_2), 7-9 (Group A_3), and 10-12 (Group A_4). The problem tasks involved mathematical proofs in a simple mathematical system. Computer-assisted instruction was used in presenting the tasks to control order of presentation and to record several aspects of subject's behavior (e.g., responses, response times, errors, and number of trials). Criterion variables used to compare groups were number of tasks solved, number of tasks attempted, number of incorrect applications of rules of inference, trials in excess of the minimum required for solution, trial difficulty, presolution time, and total time per task. Time variables and nontime variables were analyzed separately using multivariate one-way analysis of variance; both tests yielded significant differences ($p < .001$). There were significant univariate differences for number of inferences (IAR), trial difficulty (TD), and total time (TT). Using Tukey's method of multiple comparisons, the following significant results were found: $A_1 < A_4$ and $A_1 < A_3$ for TP ($p < .01$); $A_1 > A_4$ ($p < .01$), $A_1 > A_3$ ($p < .01$), and $A_1 > A_2$ ($p < .05$) for IAR; $A_1 > A_4$ and $A_1 > A_3$ for TD ($p < .01$); and $A_1 > A_4$, $A_1 > A_3$, $A_2 > A_4$ and $A_2 > A_3$ for TT ($p < .01$).

Maxwell (1975) used a block problem in her study of problem-solving performance. Three items hypothesized to be convergent in type and three items hypothesized to be divergent in type were administered to 105 students enrolled in high school geometry. On the basis of the resulting pairs of scores, these students were divided into four groups: high on both, high convergent-low divergent, high divergent-low convergent, and low on both. Subjects ($N = 49$) were chosen from each of the four groups. Each subject was observed individually while solving the Ten Block Problem (see Schwartz, 1973) which required arranging colored blocks in a four by four array. Two problem-solving trials were given to each subject. During the first trial the investigator recorded a subject's use of problem-solving processes. Next, the subject wrote a protocol describing the methods used during that trial to solve

the problem. Then the subject solved the problem a second time. The times of both trials were recorded. From these data Maxwell reported the following generalizations: (a) subjects who scored high on the divergent-in-type items made fewer generalizations in their written protocols, used trial-and-error solution methods more frequently, and took more time on the second trial of the Ten Block Problem than subjects who scored low on the divergent-in-type items; (b) trial-and-error played a major role in the problem-solving task initially and a minor role, subsequently, as the solution was approached; (c) trial-and-error increased the time needed to work the problem and seemed to be one of the main characteristics of an ineffective problem solver; and (d) girls made fewer generalizations in their written protocols, used trial-and-error more frequently, and, on the average, required more time to solve the problem during the second trial than the boys.

Correlates and Factors of Problem-solving Performance

The studies in the previous section were concerned primarily with assessments of problem-solving performance. An equally interesting topic is the search for those cognitive, affective, personality, school, and demographic variables which are related to problem-solving performance. Three studies conducted at the University of Wisconsin-Madison investigated the relation of variables in these classes to problem-solving performance. They are described in Chapters Nine, Ten, and Eleven of this volume. In addition, four other studies conducted between 1969 and 1978 will be described in this section.

Dodson (1971, 1972) attempted to describe problem-solving performance in terms of (a) mathematics achievement, (b) cognitive and personality traits, (c) teacher characteristics, and (d) school and community characteristics. The tests (Wilson, Cahen, & Begle, 1968b) from the Z-population of the National Longitudinal Study of Mathematical Abilities (NLSMA) (Romberg & Wilson, 1969) were used in this study. From the items on the mathematics achievement tests administered to the NLSMA Z-population at the end of eleventh grade, Dodson chose 40 items for his measure of mathematical problem-solving performance. Using the scores on this measure, the portion of the NLSMA Z-population who took mathematics in eleventh grade was stratified into six ability groups. Subjects (N = 1,123) for this study were a stratified random sample of 10% of the students in the Z-population for which complete test data were available. Analysis of variance and discriminant analysis were used to order the variables from best to poorest as discriminators of problem-solving performance. All of the mathematics achievement variables were significant discriminators ($p < .001$) among the six ability groups. Four variables (Z111, Z105, Z102, and Z307) were identified as the best discriminators, and three variables (Z104, Z202, and Z004) as the poorest. Dodson characterized these, respectively, as test items requiring synthesis of

relatively advanced or seemingly unrelated mathematical ideas or use of algebraic equations, and test items requiring little synthesis and involving relatively elementary mathematical ideas.

From the analysis of variance, all of the cognitive variables except one (PZ007 Picture Differences) were significantly related ($p < .001$) to problem-solving performance. In particular, the reasoning cognitive variables were better discriminators than the other cognitive variables. Generally, the personality variables were poorer discriminators between the ability groups than the cognitive variables. One of the variables (Messiness) showed no significant relation to problem solving, while its counterpart (Orderliness) had a significant negative relationship ($p < .01$). Only hypotheses were offered regarding the exploratory search of the teacher data, since data were collected from the subjects' eleventh-grade teachers and not from others who might have shaped their problem-solving performance. Finally, it was reported that the school and community variables were poor discriminators of the ability groups.

In a more limited study, Robinson (1973) tried to identify cognitive and affective characteristics of good and poor mathematical problem solvers. Initially, the following tests were administered to 115 sixth-grade students: an investigator-developed, 16-item problem-solving test; the Mandler-Sarason Test Anxiety Scale for Children (Mandler & Sarason, 1952); the Coopersmith Self-Esteem Inventory (Coopersmith, 1959); and the Kagan Matching Familiar Figures Test (Kagan & Moss, 1962). The Lorge-Thorndike intelligence scores (Lorge, Thorndike, & Hagen, 1966) and the Iowa Test of Basic Skills (Lindquist & Hieronymus, 1973) scores in reading comprehension, arithmetic concepts, and arithmetic problem solving were obtained from the school records for these students. Good problem solvers (in the top one-third on the problem-solving test) and poor problem solvers (in the bottom one-third) were compared on each of the other variables. Next, 10 good and 10 poor problem solvers of similar IQ thought aloud as they solved five mathematical problems, and their problem-solving behaviors were categorized and compared. Comparison of the problem-solving scores to the other variables showed that good problem solvers had significantly higher scores on IQ, reading comprehension, arithmetic concepts, arithmetic problem solving, and self-esteem, and significantly lower scores on test anxiety than the poor problem solvers. There was a significant relationship between problem-solving performance and reflective and impusive behavior; more impulsive students were poor problem solvers and more reflective students, good problem solvers. The probability levels of the significant results were not given in the dissertation abstract. No significant differences were reported as the result of analyzing the interview data.

The last two studies in this group are concerned with the relation of spatial ability to problem-solving performance; hence, they are akin to that

conducted by Schonberger (Chapter 10). In the first, Handler (1977) employed an experimental set of geometric spatial visualization problems to investigate the problem-solving processes and spatial visualization abilities of competent high school students. The subjects were 25 eleventh- and twelfth-grade students, each of whom participated in three individual interviews. The first interview was devoted to collecting personal data and acclimatizing subjects to the experimental procedures. The second and third interviews were used to solve the 10 experimental problems. These problems evaluated the variables of spatial visualization, imagination, visual memory, geometric concepts, and critical thinking, among others. Diagrams accompanied half of the problems. Certain exercises were alternately dictated and presented in written form to check on the interference effects of reading with visualization. Solutions to three of the problems required entirely oral responses; these were tape recorded. Student answer sheets and drawings, records of solution procedures, overt visualization behaviors, pre- and postsolution subject comments, and elapsed times were analyzed. In addition, subjects rated each problem according to difficulty, degree of confidence in their solution, and extent of their effort. The data analysis procedures for these data were not described. The processes used by the subjects were classified as deductive, insightful, or extractive. The deductive mode predominated; insightful solutions were not observed. Using the Space Relations Subtest of the Differential Aptitude Tests (DAT) as a measure of spatial visualization, good and poor visualizers were identified. Sizable discrepancies occurred in the ranks of the DAT subtests and the problem set.

Moses (1978) investigated the nature of spatial ability and spatial problems, and the roles they play in mathematical problem solving. Subjects were 145 fifth-grade students in four intact classes. All subjects were pre- and posttested using five tests of spatial ability (Punched Holes, Card Rotations, Form Board, Figure Rotations, and Cube Comparisons) (French et al., 1969a, 1969b) and an experimenter-designed problem-solving inventory. The ten problems on the problem-solving inventory represented three types of problems, namely, spatial, analytic, and equally spatial and analytic problems. Two scores, a problem-solving score and a degree of visuality score, were obtained from the problem-solving inventory. After pretesting, two of the four intact classes were randomly assigned to the 9-week experimental treatment which consisted of instruction in perceptual techniques and visual solution processes. Correlational analyses of the pretest data showed that of the five spatial tests only one, Cube Comparisons, was not correlated significantly with the others (no probability level given), spatial ability was correlated significantly with the problem-solving performance ($r = .30, p < .01$) and degree of visuality ($r = .17, p < .05$), and problem-solving performance and degree of visuality were not significantly correlated. Factor analysis of the pretest data showed that four of the spatial tests loaded on one factor while Cubes Comparison loaded on another. This result confirms the result of the corresponding

correlational analysis as does a separate analysis of electroencephalogram (EEG) data. Analysis of covariance, using the pretest scores as the covariate, was applied to the posttest data to measure the effects of the experimental treatment. There were no significant differences between the experimental and control classes on problem-solving performance or degree of visuality. The experimental treatment did significantly increase ($p < .10$) problem-solving performance on spatial problems, and there was a significant increase ($p < .10$) in spatial ability. The hypothesis that females would gain more from the treatment than males (Fennema & Sherman, 1977) was not supported.

Conclusion

This chapter has described problem-solving studies in mathematics conducted from 1969 to 1978 in the United States and Canada. The description is limited in that only studies similar to those initiated and completed at the University of Wisconsin-Madison during the same time period are included. Thus, each study meets the following two criteria: (a) the problems used in the study were mathematical problems; and (b) the research was in the area of instruction in heuristics, measurement of problem-solving performance, or correlates and factors of problem-solving ability. In these three areas 31 studies not conducted at Wisconsin were found and are described: 18 dealt with heuristic instruction, nine with assessment of problem-solving performance, and four with correlates and factors of problem-solving ability. In the following chapters nine additional studies are described. Therefore, from 1969 to 1978, 40 studies were conducted in the United States and Canada which met the two criteria imposed when searching for research reports to include in this chapter.

This author hopes, and believes, that he has found all of the research studies which meet these criteria. However, there is one extant, widely known collection of studies which is not described. Those are the Soviet studies of mathematical problem solving which have been translated and published in this country (Clarkson, 1975a, 1975b; Kantowski, 1975a; Kilpatrick & Wirszup, 1969a, 1969b, 1970, 1972; Krutetskii, 1976; and Wilson, 1975). Certainly these reports have influenced problem-solving research in mathematics conducted in this country. In fact, some of the studies described in this chapter explicitly cite their use of the techniques employed by the Soviets. Thus, one may wonder why this collection of studies is not described here. First, it seemed more important to describe, as fully as possible, the studies actually conducted in the United States between 1969 and 1978; many of the Soviet studies are older than this. Second, the Soviets' concept of the individual and individual differences, and their use of different methods of collecting, summarizing, and interpreting data limits the ability to use their findings without replicating their experiments in the United States, or to compare their results

to similar studies conducted here. Thus, it was decided not to include them in this chapter. Parenthetically, it should be pointed out that the translated Soviet studies do provide an interesting, informative perspective on problem-solving research in that country. Research should be conducted in this country on many of these questions.

In 1969 Kilpatrick (1969, 1970) commented that a good share of research in mathematics education was being done by doctoral candidates, there was an increasing number of methodological blunders, and some investigators were apparently ignorant that statistical assumptions were being violated. In addition, he stated that, because of our ignorance of mathematical problem-solving, clinical studies should be conducted in this area before large-scale, complex studies are attempted. Kilpatrick's remarks are equally true today. However, as the 31 studies described in this chapter and the nine which follow illustrate, researchers have become more aware of methodological constraints and more sophisticated in their use of statistical procedures. It also seems that Kilpatrick's advice regarding the kinds of studies that are necessary and that should be undertaken has been heeded as most of the studies between 1969 and 1978 have been clinical in nature. Perhaps the time will come when enough will be known about problem solving in mathematics to attempt larger-scaled studies.

Chapter 3

The Small Group Discovery Method: 1967-1977

Neil Davidson

Does a method of teaching mathematics exist which simultaneously fosters active learning, thinking, student pacing, and interpersonal communication? It seems apparent that the lecture method, instructional television, programmed instruction, nonprogrammed self-paced methods, the teacher-directed discovery method, and the Moore method (Whyburn, 1970) all fail in at least one of these functions.

There is no need to reiterate the familiar arguments for active learning, student thinking, and student pacing. However, the inclusion of interpersonal communication as the fourth function may surprise some readers. Interpersonal communication in education can have social benefits as well as enhance mathematical learning. Student discussion of mathematics has been emphasized by Buck (1962, p. 563):

> . . . Let me remind you that *student-student* interactions are also important in learning, and that at the professional level, much mathematical research springs from discussions *between* mathematicians. Moreover, a test of understanding is often the ability to communicate it to others; and this act itself is often the final and most crucial step in the learning process.

On philosophic, psychological, and biological grounds, various authors have stressed the affiliative needs of human beings and the social impetus for human activity (Dewey, 1916; Montagu, 1966). However, there are a number of forces in society and in education which ignore these affiliative needs and generate depersonalization, anonymity, loneliness, anxiety, and alienation (Association for Supervision and Curriculum Development, 1967; May, 1953; Sarnoff, 1966). Such forces are present in many modern universities where many students spend a substantial amount of time as anonymous members of mass lecture sections. Interpersonal communication should be emphasized because it promotes student discussion of mathematics, counters societal forces toward loneliness and anxiety, and provides personal support in the educational process.

It appears that these goals can be achieved by dividing the class into small groups where the students can discuss mathematical problems with a few colleagues. The number of students per group is deliberately restricted to increase possibilities for personal contact. Small group instruction can foster active participation and, to a large extent, student pacing. Moreover, in small

group instruction, the amounts of discovery and guidance can be varied, depending on the desired level of student thinking. Finally, small group learning in conjunction with discovery learning offers possibilities for curriculum development if differences in student learning are observed.

In this study, the subject area of elementary calculus was selected for exploration. Positive results in calculus instruction were attained previously through the "student experience-discovery approach" of Cummins (1960) and the heuristic problem-solving approach of Larsen (1961). The Moore method was used by its developer to teach calculus (see Moise, 1965). Kingsbury (1963) made successful use of a self-pacing activity method in calculus instruction, and Turner, Alders, Hatfield, Croy, and Sigrist (1966) used small group instruction as a supplement to several large lectures in calculus per week. There was no record of a previous attempt to develop a method of calculus instruction combining discovery learning with a small group approach. This combination, named the "small group discovery method," was first employed by the author in a 1967-68 pilot study with a freshman calculus class at the University of Wisconsin-Madison.

The remainder of this chapter is divided into the following sections: the design of the small group discovery method using Dewey's educational philosophy, an elaboration of the design based on studies in social psychology, a description of the classroom social climate during the pilot study, a description of how students interacted with the mathematics content, data for evaluating the pilot study, conclusions and questions for investigation, and work completed since the pilot study.

Classroom Practices Derived from Educational Philosophy

George Polya (1965) has emphasized student thinking, active learning, discovery learning, and interest in mathematics. It has not been widely recognized in the mathematical community that these are particular aspects of a general philosophy of education and of life, whose foremost advocate in education was John Dewey (1916, 1938). The small group discovery method was designed in accordance with that philosophy. Supporting evidence and further elaboration was provided by studies in social psychology. A description of Dewey's philosophy is beyond the scope of this chapter. However, it does summarize classroom practices derived from Dewey's philosophy and applied in the 1967-68 pilot study with a freshman calculus class.

During the pilot study, students learned mathematics by doing mathematics. The approach was one of guided discovery in which mathematical topics were introduced as questions to be investigated by the students. The students, with limited guidance from the teacher, formulated definitions, stated

theorems, proved the theorems, constructed examples and counterexamples, and developed techniques for solving classes of problems.

The classroom activity was a social process taking place in small groups. Within each group there was to be a cooperative atmosphere where students worked together to solve the problems.

The teacher adopted a democratic leadership style by participating in the students' activities, but not in a highly directive way. The teacher spent most of each class period with the small work groups. He kept track of the progress of the groups, made corrections and suggestions, clarified notation, gave hints, checked solutions, provided encouragement, and tried to see that the groups functioned smoothly.

At times the teacher talked with the entire class, generally for no more than 5 or 10 minutes each day. During these brief discussions, new concepts were presented, questions were raised and answered, problems were assigned, hints were given, clarifications and summaries of student work were made, and student discussions were moderated.

There was no textbook for the experimental course. To guarantee that all of the major calculus topics were included, the teacher proposed problems; this practice departed from Dewey's philosophy. The presentation of mathematical topics proceeded from the more concrete to the more abstract. The discovery of new ideas was emphasized rather than the expression of ideas in an impeccable form. Professional standards of rigor were not imposed upon these students, and initial development of ideas was informal in character. It was anticipated that the need for increased precision and theoretical security would become apparent to the students as they handled more difficult or abstract problems.

Since interest in the mathematical content was intended to provide the major source of motivation for the students, the teacher attempted to determine which topics were of intrinsic value, which appeared to be useful (instrumental), and which had little value from the students' perspective. In addition, a nonthreatening classroom atmosphere and reduced emphasis upon grades (the use of an A-B grading scale, elimination of in-class examinations, and student determination of some grading policies) was employed.

Skills were developed under conditions where thought was necessary. The students developed the techniques for solving each class of problems presented to them. Additional practice occurred when problems differed from each other and when judgments were needed to find solutions. Whenever possible, skills were attained by solving problems of instrinsic value to the students.

Within this basic framework many questions occurred to the students. Investigation of student-generated questions was one of the important class

activities. Daily notes, prepared by the teacher, contained a record of problems solved by the students and of questions which arose.

Classroom Practices Derived from Social Psychology

Supporting evidence and elaboration for these classroom practices were provided by empirical studies in social psychology and group dynamics. In a classic study of leadership styles conducted by White and Lippitt (1960), the adult leaders of boys' clubs were trained to be proficient in using authoritarian, democratic, and laissez-faire styles of leadership. Characteristics used in defining and comparing the styles included the degree of involvement of the leader with the group, the degree of warmth or impersonality of the leader, the locus of control and procedures for decision making, the use of orders versus suggestions, and the objectivity and frequency of evaluative comments by the leader.

The authoritarian (autocratic) leader determined all policies, dictated techniques and activities one step at a time, dictated work tasks and work companions, offered much nonobjective praise and criticism, and gave orders and disrupting commands.

The democratic leader helped the group make policies through group discussion and decision making, provided an activity perspective and sketched general steps toward the group goal, suggested alternative procedures from which group members could choose, allowed members to select work partners and to determine division of labor, offered a small amount of objective praise and criticism, provided guiding suggestions when needed, and acted in a friendly and equal manner.

The laissez-faire leader allowed complete freedom for group or individual decisions, participated to a minimal extent, supplied materials, supplied information upon request, commented infrequently upon members' activities, and offered almost no appraisal of the work.

White and Lippitt found that the quality and quantity of the work was greater in the democratic situation than in the laissez-faire situation. There was not a clear distinction in terms of quantity and quality of work between the authoritarian (autocratic) and democratic situations. However, genuine interest in the task was "unquestionably higher" in democracy than in autocracy.

There were numerous indications that morale was higher in the democratic situation than in the autocratic situation. The autocratic groups were marked by discontent and a tendency toward group fragmentation. In an aggressive reaction to autocracy, there was a large amount of hostility; in a sub-

missive reaction, the group atmosphere was subdued and low-spirited. In both reactions there were submissive and dependent actions toward the adult leader.

Anderson (1963) reviewed 49 empirical studies defining leadership along an authoritarian-democratic dimension. Most studies did not include the laissez-faire style. Thirty-two of the studies deal with leadership in educational settings; the results were not conclusive with respect to measures of student learning. However, morale was generally higher in the democratic (learner-centered) groups, except in a few cases involving "high anxiety about grades which are awarded on the basis of final examination scores . . . ."

On the basis of the research by White and Lippitt (1960) and by Anderson (1963), the teacher of the calculus class in the present study used a carefully specified style of democratic leadership. He provided a perspective on each day's mathematical activities in a brief discussion with the entire class and spent most of the period working with small groups. He refrained from giving orders or disrupting commands. There was only a minimal amount of objective, constructive praise and of criticism. Usually criticism was directed to the work group as a whole and not to individuals. The teacher offered guiding suggestions at times when they were needed; these included mathematical hints and suggestions about work organization and group functioning. He sometimes provided technical information upon request, and stimulated self-direction by encouraging members to detect group errors and think through and elaborate upon their ideas. The teacher developed a friendly, social relationship with the students and behaved in an egalitarian manner which included reciprocal use of first names. Finally, many policies in the calculus class were arrived at through group discussion and decision-making by a majority vote.

In this study, decisions about cooperation or competition within the work groups were made by the teacher on the basis of research conducted by Deutsch (1960) at Massachusetts Institute of Technology. Deutsch found that the productivity of a competitive discussion group was reduced by poor coordination, duplication of efforts, inattentiveness to the ideas of others, obstructive and self-defensive behavior, and group conflict. In a cooperative situation, as compared with a competitive situation, the group members were more friendly, listened more attentively, and understood the ideas of others better. Moreover, the group discussion was more productive in terms of the quantity and quality of problem-solving ideas generated. In accordance with these results, the teacher in the calculus class promoted cooperation within each work group by checking the group solution without asking who was responsible for it and by not giving individual grades for classwork. He emphasized the need for joint efforts to solve difficult problems, the importance of listening carefully and building upon the ideas of others, the fact that one

person's good idea helps the entire group, and the goal of solving the problem so that all members understand the group solution.

Studies have shown that pressure to "go along with" a group can lead to the modification or distortion of individual judgment or perception (Asch, 1960). Hence, this conformity pressure and independent thinking are antithetical to one another. Fortunately, it is possible to reduce conformity in problem solving by developing group standards which encourage members to follow their own judgment (Deutsch & Gerard, 1960). The teacher in the calculus class developed such standards by emphasizing the importance of independent judgment, the legitimacy of disagreement, and the obligation of group members to give reasons supporting their statements. The teacher intervened as a mediator when students looked puzzled or confused or when several group members put undue pressure on a dissenter. The teacher emphasized the distinction between thoughtless conformity and a change of opinion based upon a thoroughly understood argument.

A commonly held misconception is that every group must have a leader (Cartwright & Zander, 1960). In the calculus class there was no clear need to appoint a leader for each group. Moreover, there were risks involved in designating group leaders, since the opportunity for active participation by the followers would then be reduced and since there might be hostility between the leader and those who wished to depose him or her. Therefore, the work groups operated without designated leaders. Although it was not possible to create a completely egalitarian work group, it was possible to place limitations upon the discrepancy in power between the most active and least active group members. No person was allowed to dominate the discussion in a manner that excluded or severely limited contributions from others. Whenever necessary, the teacher influenced the dynamics of particular groups by drawing certain members into the discussion, suggesting that different people assume primary responsibility for writing problem solutions on the blackboard, and using other techniques to promote cooperation.

It was necessary to maintain small work groups, since the opportunity for active participation would decrease as group size increased. There was some empirical evidence available of the effects of group size on group interaction in nonmathematical discussions. Two-person discussion groups were found to be marked by a tense, cautious atmosphere in which the members tended to avoid conflict and expression of their ideas. In two-person groups there was no one to resolve differences, and either member could bring the group to a halt by disagreeing or withdrawing (Bales & Borgatta, 1961). Three-person groups tended to break up into a pair and an isolated member. Four-person groups could split into two subgroups of equal size and thereby produce a protracted argument or deadlock (Bales & Borgatta, 1961; Mills, 1960). Five-person groups entailed the dangers of competition, exclusion of members from the discussion, and the need for a definite leader (Slater, 1958).

The experimental evidence was sufficient to rule out the two-member group and the five-member group in an effort to achieve active student participation. There was no clear case for selecting either the three-member group or the four-member group, so the teacher simply chose the four-member group for the pilot study. It was not clear that the teacher could give adequate attention to more than three groups during his first trial of the instructional method, so the class for the pilot study was limited to 12 members.

Kilpatrick (1969) and Symonds (1958) have reviewed studies of the detrimental effects of grades and anxiety problem-solving performance and creativity. Moreover, the grading problem has presented "the greatest obstacle" to the success of some past attempts at student-centered instruction (McKeachie, 1954). These observations provided support for the philosophically based decision to use a permissive grading scheme for the calculus class.

The Classroom Social Climate During the Study

For the first trial of the small group discovery method, the teacher set some entrance requirements for the students. In order to join the class, a student had to be a freshman or sophomore with little or no prior knowledge of calculus, have grades of A or B in high school mathematics, and be at least mildly interested in mathematics. Students were selected for the class through interviews with the teacher at the course assignment committee. Only one student who was interviewed decided not to join the class. The pilot class consisted of four female and eight male students; there were 11 freshman and one sophomore.

Leadership by the Teacher

The teacher made use of a democratic style of leadership during the pilot study. Many policies in the class were determined through group discussion and decision making by a majority vote, with the teacher serving as discussion moderator. The students decided the membership and division of labor in their groups and the time schedule for changing membership. They selected take-home exams from a set of 11 alternative grading policies, permitted each student to begin work on the exam at a convenient time during a one-week period, and decided not to make up definitions or terminology which would conflict with standard mathematical usage.

The teacher gave a perspective on each day's mathematical activities in a brief discussion with the entire class. He often introduced new topics in the form of questions for investigation by the students, such as: "How can we find the area under a curve?" 'What happens at a high or low point on a curve?" "What can you say about a function which vanishes at the end points of its interval of definition?" "Can we find a formula for the derivative of a product?" "How can we find the volume of the solid obtained by revolving a curve around an axis?"

Almost all mathematical discussions with the entire class lasted for less than 10 minutes. The discussions set the stage for the main activity of small group problem solving. Just enough input was provided in discussions so that the groups could function productively for the rest of the class period.

The teacher usually used praise and criticism in an objective, constructive manner and directed it to the whole work group. However, there were two examples of personal criticism during the year. In one situation, the teacher said to someone, " . . . ., you still don't know your derivative formulas." She immediately froze up and was less friendly to the teacher for the next week, during which she participated less than usual. In the other situation, the teacher said to a student, "Why don't you look at the solution and point out your mistake." The sarcastic reply was, "Well now, if I'd known it was a mistake I wouldn't have done it, would I?"

The teacher almost never gave orders or disrupting commands. In one exception to this, the teacher stopped the groups in the middle of proving the chain rule and told them to think about it overnight. While some students were relieved, others expressed resentment; "I sure hate to get cut off in the middle of a problem." On several occasions the teacher asked the groups to stop working near the end of the period so that he could present a summary. It quickly became apparent that students did not care to have a summary at the end of the period. They kept on talking about the problems, and several students stated that they knew what they had done and required no further reiteration to understand it. The practice of end-of-period summaries was rapidly abandoned.

The teacher found it easy to keep track of group progress, since students wrote their problem solutions on the board. He frequently did not wait for a request for assistance, but offered suggestions at times when they appeared to be needed. Usually, a visit with a particular group took less than 1 minute. Sometimes a visit lasted only 10 or 15 seconds—for example, if it was only necessary to point out an arithmetic mistake, ask the reason for a step, or check a simple solution. However, on difficult proofs the groups needed considerable assistance, and visits to groups lasted 2 or 3 minutes. If the teacher stayed too long with one group, members of other groups began calling for help.

Guiding suggestions of a mathematical nature were given in the form of hints, sometimes using the heuristic techniques of Polya (1965). Here are some examples:

1. The teacher frequently asked the students to concentrate on the given data, the desired result, and relationships between the two. This helped in many proofs, especially those using the definitions of the limit.

2. The teacher sometimes suggested that groups attempt to use prior results and to reason by analogy. For example, when students had trouble

deciding whether a certain function had no limit or two limits at a point, the teacher suggested a comparison with sequences such as 2, -2, 2, -2, . . . , where the same issue had been settled previously.

3. It was occasionally helpful to suggest that students consider a simple instance of a general problem. This was done with $n = 2$ and $n = 3$ in guessing the formula for the derivative of a product of n functions.

4. General results were sometimes formulated by considering special cases. The students correctly surmised the fundamental theorem of calculus after computing

$$\int_a^b x^k dx,\ k = 1, 2, 3.$$

5. It was sometimes useful to suggest that the students discover or confirm results by drawing pictures. This was suggested when students could not remember if $d(\sin x)/dx$ is $\cos x$ or $-\cos x$.

6. The suggestion to guess the answer to a problem sometimes led to some surprises. Students were convinced that the derivative of a product should turn out to be the product of the derivatives.

7. A slight shift in notation occasionally made a big difference in problem solving. In their first encounter with implicit differentiation, students had great difficulty in finding dy/dx *for* $x^2 + y^2 = 1$. The hint to replace y by $f(x)$ readily enabled students to find $f'(x)$.

8. There were occasions when a hint given only once lasted for the remainder of the year. In the proof of the formula for the derivative of a product, the hint was given to add and subtract the same term. For the rest of the year the students correctly used this technique when needed.

The teacher sometimes offered guiding suggestions with respect to the work organization and functioning of a particular group. Students often wrote four or five attempted solutions all over the board, and no one could tell where one idea ended and the next began. Many students omitted key symbols — for example, writing $\sin x = \cos x$ instead of $d(\sin x)/dx = \cos x$ or $\int \cos x\, dx = \sin x + c$. This caused great confusion on complicated problems, and suggestions about blackboard technique were much needed. Suggestions about the social functioning of the groups are described later.

The teacher provided technical information on request if the development or recall of that information was not a key part of the problem at hand. For example, a request was always honored for an approximation of the number e to five decimal places. A request was never honored to provide the formula for $d((f(x)^n)/dx$. Other items of information, for example an identity for $\cos 2\theta$, were provided for some problems but not others.

The teacher checked the group solutions of all the difficult problems or theorems. In other problems, checking preferences varied. Some group mem-

bers always wanted their solution checked; other group members were quite confident and erased their solutions without teacher checking. When enough board space was available some groups left one solution up for checking while working on another problem.

The teacher attempted to stimulate self-direction by encouraging people to look for errors in their group's solutions. Many errors were caught by the students themselves, and others were detected by the teacher. There were computation errors, incorrect applications of basic formulas [$d(\sin 3x)/dx = \cos 3x$], errors in basic algebraic facts, logical errors of many types (e.g., circular reasoning and proving a conclusion without using the hypothesis), errors for over-generalization [$d(e^x)/dx = x\,e^{x-1}$], and errors of notation [if $f(x) = x^3$, then $f'(x^3) = 3x^2$]. The teacher was surprised by the students' frequent shifts from error to insight.

Although all groups began each new topic on the same day, some groups moved more quickly than others. The teacher always had some challenging extra problems for groups which finished early.

The teacher developed a friendly relationship with the students and tried to reduce the gap in status between the students and himself. He often arrived a few minutes before class to chat about campus events, world happenings, and so on, but saved personal problems for discussion outside of class. He suggested a mutual first name basis, which made some students uncomfortable at the start of the year: "Are we supposed to call you Professor Neil or Dr. Neil, Mister Neil, or what?" This issue dissolved after a few weeks.

Dittoed notes were prepared by the teacher after each class meeting to record the students' accomplishments on that day. The notes were distributed at the following class meeting. Although the teacher had expectations about the material to be covered on any given day, these expectations were wrong more often than not. Students frequently encountered unexpected difficulties, came up with novel problem solutions, or pursued questions not planned by the teacher.

Group Formation and Size

The students decided to change groups every 2 or 3 weeks or at the end of major units of content. The process of changing groups was awkward, though brief. Students had different styles of coping with the change process. Some said directly, "Let's work together." One girl always went straight to "her corner" and waited for others to join her. Some students sat and pretended to do homework until the groups were basically formed; then they looked around for a vacancy. One person sometimes wandered around the room looking lost until settling upon a group. Despite this awkward process,

the students refused a suggestion by the teacher to form groups by writing down confidentially their most preferred and least preferred group members.

During the pilot study there were four members in each work group. The four-member group functioned well on theoretical problems, but was sometimes too large for optimal practice with standard computation problems. In computational problems the groups often split spontaneously into two pairs. When one member was absent the remaining three functioned well except when the problems were very difficult or the missing member normally exercised much leadership. On rare occasions when two group members were absent, the remaining pair either limped along or split up and joined different groups for the day.

Cooperation

The teacher fostered cooperation and discouraged competition by talking with group members. Here are some examples of teacher comments made in various situations: "Some of these problems are very hard and you have to work together to solve them quickly." "There is no need to blame anyone for a mistake." "How about listening? Are you really disagreeing or just saying the same thing in different words?" "Is it possible that you're both right?" "The group goal is not only to solve the problem but to do so in a way that everybody understands."

Roughly two-thirds of the students were cooperative, but at least three individuals believed in and practiced competition. When two particular competitors were in the same group they argued intensely, tended to exclude the other two members from the discussion, and were impatient in answering their questions. When this pattern became clear, the teacher asked the two competitors to work in different groups. On many occasions, a cooperative group producing a small number of ideas solved problems as quickly as a competitive group which generated many ideas but could not agree upon them.

The work groups usually functioned as separate units with little communication between them. Members of one group almost never borrowed ideas from the solutions of other groups. On numbered lists of problems, people sometimes compared the problem number, but not the solution, between two groups. Competition between groups occurred spontaneously, but very rarely, and only on numbered lists of rather easy problems. When competition arose between groups, it was done in the spirit of a lively game which no one took seriously.

Leadership by Students

Although the groups operated without designated leaders, some group members were much more active and influential than others. For the entire year, one girl emerged quite clearly as the task leader on most problems no matter who else was in her group. No one else was able to match her consistent quickness and enormous enthusiasm. There were interaction difficulties in

some of her groups at the beginning of the year, but these disappeared as people deliberately chose to work with her or to avoid her group. At the other extreme was a boy who was always the least active and the least influential member of his group. He rarely made any mathematical suggestions or wrote any solutions on the board. However, when he did contribute an idea or answer a question, he was almost always correct.

The behavior of the other group members fell between these extremes. More than half of the students were very active particpants in problem solving througout the year. Some individuals participated to a greater or lesser extent, depending on who else was in their group. Sometimes, a less influential group member had the basic idea for a solution, but the idea was either not heard or not accepted. When the teacher came over and made the very same suggestion, the person with the idea blurted out: "See, that's what I was trying to tell you 5 minutes ago!"

Students were very reluctant to criticize dominant behavior, even when it clearly interfered with group progress. The system of taking turns to write down the problem solution helped to some extent, especialy on numbered lists of problems. Another idea was for different group members to become "experts" on certain problems, solve them outside of class, and present them to their group. This approach turned out to be a disaster and was quickly abandoned.

Conformity

Conformity pressure in the groups had to be reckoned with, but was not a serious difficulty. For example, on various occasions the teacher heard someone say, "I suppose we've solved the problem but I don't see why it works." Another student replied, "Never mind why it works, it just does. Let's go on." A quick teacher intervention to check understanding resolved the issue. There were only a few incidents in which three group members put pressure on a single dissenter. In one case there was an intense discussion of the relative merits of approximating areas by rectangles or by little squares. The dissenter, who favored squares, became visibly upset. The teacher then intervened, pointing out that both sides were right for different purposes, and the approach with squares would be used in a later semester.

The Interaction Between the Students and the Mathematics Content

Through daily conversations with the students and observations of their work, the teacher learned much about their reactions to the subject matter. The students developed problem-solving techniques and proved the major

theorems as expected but affective and cognitive aspects of the student interaction with the content were sometimes surprising. The close contact with students in the study groups helped the teacher gain much insight into student perceptions of calculus.

During the pilot study, most students had difficulty with testing universal statements by particular instances, recalling definitions, and transferring information from one problem to another. The students frequently did not test incorrect universally quantified statements by using specific instances. Examples of this included the incorrect formulas $1 + \sec^2 \theta = \tan^2 \theta$, $\cos(x + y) = (\cos x)(\cos y) + (\sin x)(\sin y)$, $d(\sec x)/dx = \tan^2 x$. Each time the students wrote down such an incorrect identity, the teacher asked if there was any easy way to test the truth of their statement. It was usually necessary to tell the students to try their statement with a particular value of the variable.

The students were apparently not used to thinking in terms of definitions, and they tended to forget major definitions from one day to the next or even from one problem to the next. Most students were persistently unable to recall definitions of the limit, the definite integral, and continuity. The definition of the derivative fared somewhat better than the other definitions, perhaps because it was a simple formula which was frequently used. The students often tended not to use the major definitions, even in problems where use of the definition provided the only possible approach. For example, the students did not think of using the definition of the integral to test the integrability of the following function: $f(x) = 0$ if x is rational, $f(x) = 1$ if x is irrational, where $0 \leq x \leq 1$.

There was a noticeable tendency for the students to treat all problems as separate and unrelated entities. For instance, the groups first evaluated $\int dx/(a^2 + x^2)$ by means of the substitution $x = a \tan \theta$. Instead of using this result, the groups then evaluated $\int dx/(9 + x^2)$ by the substitution $x = 3 \tan \theta$. Although some repetition can be a useful aid in learning, many students repeated the same useful work over and over again, long after they had mastered the appropriate integration technique.

In working with derivatives of composite functions, most students did not perceive the need to apply the chain rule in new situations. Although the student groups correctly developed the formula for the derivative of each major new function, they then made erroneous statements such as $d(\sin 2x)/dx = \cos 2x$, $d(e^{3x})/dx = e^{3x}$, $d(\ln 4x)/dx = 1/4x$. In each instance it was necessary for the teacher to remind students that they were dealing with composite function.

In problems of integration almost all students had persistent difficulties in working with the differential. For example, in evaluating $\int (\sin^3 x)(\cos x)\, dx$, the groups let $u = \sin x$ and evaluated $\int (u^3)(\cos x)\, dx$ without expressing all of the integrand in one variable. Later the groups used

the relation of $u = \sin x$ to derive the incorrect expressions $du = \cos x$, or $du/dx = (\cos x)\, dx$. In many problems with integration by substitution or by parts, the students first forgot to convert all the variables in the integrand. Then, when they tried to do so, they frequently ended up with either too many or too few differential symbols.

In problems which could be solved in several different ways, the students often preferred to use the technique they had learned first. For example, the integration technique of trigonometric substitution was first introduced by the problem of computing the area of a circle. To evaluate

$$\int_{-r}^{r} \sqrt{r^2 - x^2}\, dx$$

the groups used polar coordinates and set $x = r \cos \theta$. Then, in many other integrals involving the expressions $r^2 - x^2$, the students always used the substitution $x = r \cos \theta$, rather than the more standard substitution $x = r \sin \theta$. Several students said that $x = r \sin \theta$ would work, but that they liked their first approach better.

The students' intuitive notions about sequences were surprising to the author. Almost all the students believed initially that the listing 1, 1, 1, 1, . . . did not describe a sequence, since the n-th term did not change and was not specified by a formula involving n. After resolving this issue, almost all students stated that the sequence 1, 1, 1, 1, . . . did not have a limit, since "it's not getting close to any number; it's there already."

Most students stated that the sequence, 0, 2, 0, 2, 0, 2, . . . converged to two limits, and were upset when the teacher said the sequence had no limit. Their discomfort was alleviated somewhat when the teacher introduced the notion of a subsequence.

In trying to solve problems or do proofs using the definition of the limit of a sequence or a function, the students encountered great conceptual and technical difficulties. Comments from several students indicated that they did not perceive the statement as a reasonable definition. "If that's a definition, it's the weirdest one I've seen in my entire life." Moreover, most students did not find the proofs of limit theorems useful. "There's no reason to prove a theorem unless there is some doubt about the result, and I never had any doubt about the sum of the limits being the same as the limit of the sum." Many students were not convinced by proofs of the limit theorems. "That proof is nothing but a bunch of equivalent statements with complicated notation. It doesn't prove anything to me." These attitudes and difficulties were not caused by lack of prior concrete experience; the students had spent several weeks working with a variety of sequences before encountering the formal definition of the limit.

The student concept of a function seemed to include several basic but unstated assumptions. Students invariably drew the graph of a function as a smooth curve with a small number of relative maxima or minima. A student

said, and others agreed, that "there are only three possibilities at an endpoint of an interval. Either the curve comes in level or it comes from below or from above." It appeared that the student concept of a function on a closed interval actually meant a continuous, differentiable function with a finite number of maxima and minima.

Students were almost always unable to state or recognize the definition of continuity. In many problems they automatically used the property

$$\lim_{x \to x_0} f(x) = f(x_0)$$

without asking whether the property held for the given function. Morever, students tended to assume the existence of absolute maximum and minimum values for any function defined on a closed interval. Most believed that there was something unnatural or artificial about functions with discontinuities. As they put it, these functions were "made up" by moving points out of their proper location, adding points which did not belong in the domain, putting in steps, or creating infinitely many oscillations in the graph.

The students seemed to think at times that all functions were differentiable. For example, for the function $f(x) = |x|$, the students stated that they were going to find $f'(0)$. When the right- and left-hand limits of the difference quotients turned out to be different, most students thought they had made an arithmetic mistake. Moreover, the students almost always reversed the relationsip between differentiability and continuity and stated that all continuous functions were differentiable.

Many students made a distinction between theory and problems. As one student put it, "Calculus should be 25% theory and 75% problems." The distinction between theory and problems seemed to depend largely on the presence or absence of arbitrary functions. Although most students preferred problems over theory, they sometimes distinguished between useless theory and useful theory. Useless theory consisted of propositions intended to "prove the obvious" or "straighten out things we already know." Most students deemed as useless the definition of the limit and the development of the natural logarithm as an integral. Useful theory consisted of general propositions which had applications to interesting problems with specific functions. Many students accepted as useful theory the proof of the fundamental theorem of calculus and the development of the formula for the volume of a surface of revolution.

The difficulties encountered by better-than-average students in a discovery approach to calculus were quite surprising, even to an experienced teacher of calculus. Hopefully, these difficulties will not obscure the success of student groups in proving the major theorems of calculus, developing techniques for solving classes of problems, stating insightful conjectures, and coming up with problem solutions and proofs not previously known to the teachers.

Data for Evaluating the Pilot Study

Comparison between Two Methods of Instruction

As part of the evaluation for the pilot study, a final examination was given to students in the small group discovery class, or "discovery group" and to a control group. The control group consisted of 51 students who learned calculus via the lecture-discussion system. The 51 students were a subset of a lecture class which met with a professor for three lectures per week. On the other 2 days, the 51 students met in four separate discussion sections led by four different teaching assistants.

There were differences between the two groups with respect to group composition and conditions pertaining to the examination. The 12 members of the discovery group were all volunteers for a special class; all of these students had grades of A or B in their high school mathematics classes. The students in the control group were not volunteers and had not been subject to any special entrance requirements. In comparing the two groups, there was no attempt to control such variables as SAT scores, IQ scores, sex, or grades in high school mathematics.

The final examination was administered at the end of the second semester of the 1-year pilot study. During that year, the students in the discovery group had taken no examinations or quizzes in class. Students in the control group had taken a final examination during the first semester, several hourly examinations during both semesters, and quizzes at the discretion of the various teaching assistants.

The discovery group took the final examination designed for the control group. The examination involved the recall of facts and standard computation skills; it did not require the solution of any difficult problems or the formulation or proof of theorems. The examination did not include material such as limits of sequences which was covered in the discovery group but not in the control group. However, the 25 items on the examination did include seven items covered in the control group but not in the discovery group; the members of the discovery group were told to omit these items. Thus, the members of the discovery group had more time available during the 1-hour test. The comparison was made on the basis of the 18 items common to both groups; the examinations of the discovery group were scored by the author.

The raw scores, mean, median, and standard deviation on the final examination are presented in Table 1. The mean and median for the discovery group were 55.25 and 55.5, respectively. The mean and median for the control group were 52.35 and 53.0. On a section-by-section basis, the mean and median were higher for the discovery group than for the four sections in the control group, with the exception of one section in which the median was the same as for the discovery group. In addition, the standard deviation was 7.59 for the discovery group and 11.57 for the control group. An F-test was run on data,

Table 1
Scores on the Final Examination

Group	Discovery	Control 1	Control 2	Control 3	Control 4	All control groups
Number of students	12	12	16	12	11	51
Total of scores	663	662	864	601	543	2,670
Arithmetic mean	55.25	55.17	54.00	50.08	49.36	52.35
Median	55.5	55.5	50.5	51.5	51.0	53.0
Standard deviation	7.59	7.00	11.86	14.66	11.77	11.57

yielding an F-ratio of .679 ($p < .413$). Hence, the difference was not statistically significant.

The author tabulated the number of perfect solutions achieved in each group for each item of the test, excluding those items which were omitted for the discovery group. From these data the average number of perfect solutions per student in each group was computed. The average number of perfect solutions per student in the discovery group was 12.50. The corresponding numbers for the four control sections and the total control group were 12.25, 12.38, 10.92, 10.91, and 11.69, respectively. Therefore, the average number of perfect solutions per student was higher in the discovery group than in each control session taken separately and in the total control group.

Take-home Examinations

In the discovery group there were seven take-home examinations during the pilot study — three during the first semester and four during the second semester. The results of these examinations, with scores converted to a 100-point scale, are summarized in Table 2.

On five of the seven take-home examinations, the means and medians were higher than 80. On the remaining two examinations, which were quite difficult, the means and medians were higher than 70. Among the 84 individual scores on the examinations, only 10 scores were lower than 70. The absolute minimum score was 60, which occurred only once and on the first exam. No member of the class was the low scorer on more than two of the seven examinations.

Table 2
Results of the Take-home Examinations

Exam number	1	2	3	4	5	6	7
Mean	72.6	87.0	72.6	80.8	86.6	88.8	90.5
Median	74.0	86.5	72.0	82.0	86.5	91.0	90.5
Standard deviation	7.90	7.97	6.30	9.16	7.53	6.38	8.92

The Questionnaire

A 90-item open-ended questionnaire was used to determine student reactions to the small group discovery class. In constructing the questionnaire, it was decided that open-ended items might provide a good deal of information not readily available in another form, although such items would be difficult to classify and count. The questionnaire was loosely constructed and no claim was made about its reliability; it was not suitable for use in a carefully controlled investigation. Nevertheless, for an informal evaluation of the first trial of an instructional method, it furnished the necessary information.

The questionnaire was divided into five major sections, dealing with the work groups, the mathematics content, the teacher, various practices and policies, and basic reactions to the class. The questionnaire was given to the students about 2 weeks before the end of the semester. For almost every item on the questionnaire, the student responses were classified into various categories and counted, although for certain items it was not possible to form very neat categories. Since the flavor of the student responses cannot be conveyed by an item summary, there follow some quoted responses to one question.

How did working with other students influence your learning?

Categorized Responses	*Frequency*
a. Uncertain	2
b. Positive Effects	10

Sample Responses

a. It is very difficult to judge.

b. Other students, no matter who, force you to learn more.

It helped me because I gained confidence showing people how to work a problem, also realized my limitations when people showed me how to do a problem.

A lot of times when I did not understand something the other members of the group helped to clear things up.

The working out of problems together not only removed much of the frustration of working difficult ones by oneself, but it also helped keep up a constant renewal of interest.

I learned to depend on working it out myself or with the help of others instead of relying on a book. In other words I think I developed a little original thinking.

I think I learned a lot more this year than I did in all 3 years of high school math.

I think you learn from students while you're taught by teachers. I think you know what I mean. With a student you understand, with a teacher you too willingly accept.

Summary of Major Results from the Questionnaire

- Two-thirds of the class members either did not enjoy the theory or enjoyed it only sometimes.
- No student reported a decrease in interest in mathematics during the pilot study, and one-third of the students reported an increase in their interest in mathematics.
- No student reported a decrease in skill in problem solving during the pilot study, and more than half of the class members believed that there was an increase in their problem solving skill.
- Most of the students said that working with others had positive effects upon their own learning.
- Only one-fourth of the class members were never concerned about covering enough material during the pilot study.
- Most of the students reported that the teacher spent the right amount of time with each work group and gave enough hints.
- More than half of the students said that the teacher was effective in giving hints, and the other students said he was sometimes effective.
- Everyone perceived the teacher more as a helper and guide than as an evaluator and critic.
- Almost all the students reported a closer, more personal relationship with their mathematics teacher than with their other teachers.
- All the students believed that the two-member group was too small and the five-member group was too big. There was a division of opinion about the relative merits of the three-person group versus the four-person group.
- Three-fourths of the class members expressed a desire to avoid working with various individuals.
- More than half of the class members were sometimes in a group with a person they didn't like.

- Two-thirds of the class members reported feeling completely free to ask questions when they didn't understand something. The other people felt very free, but this depended on the person being asked.

- Three-fourths of the class members said that they had an adequate opportunity to express their ideas in their groups. The other people said it depended upon the particular group.

- Almost half of the class members reported competing with others.

- While some of the groups functioned very well, others did not.

- Most of the class members said that working problems every day did not become routine or monotonous.

- Three-fourths of the students said that they never read a calculus book during the pilot study.

- Three-fourths of the students reported feeling little or much less grading pressure than in their other classes.

- Half of the class members saw other class members socially.

- Most of the students said that their calculus class was better, more stimulating, or much more stimulating than their other classes.

- More than half of the class members said that their attitude toward the class changed for the better during the year, and the other students said there was no change in their attitude.

- Seven students said that they would have no reservations about taking future courses taught by this method. Five people expressed reservations, which were the following: doubt about learning as much as in ordinary classes, the great dependence of this method upon the particular instructor, the desire to try a lecture-quiz course in math, the wish to avoid contact with other people, and fear of not being able to make it in a regular class.

- The students mentioned the following advantages of the method: less grade pressure, more interesting, easier, more opportunity to clear up questions, more fun, more challenging, and greater student-teacher contact. In addition, it builds good relationships with others and stimulates desire to learn math, gain ideas from other people, learn to teach, get more help, learn more thoroughly, develop greater understanding, think for oneself, and think creatively.

Conclusions and Suggestions for Investigation

The pilot study demonstrated that an entire first-year course in calculus can be taught by the small group discovery method. The student groups succeeded in proving the theorems of calculus and developing tech-

niques for solving various classes of problems with only limited guidance by the teacher.

An informal comparison, not employing a formal experimental design, was made of student achievement in the small group discovery class and in a lecture-discussion section. On the common items of a final examination dealing with basic facts and computational skills, the small group discovery class performed at least as well as the lecture discussion class. Inspection of the syllabi for the two classes showed that the small group discovery class dealt with as much material as the lecture section, but not with exactly the same material. Finally, the students in the discovery class performed very credibly on seven nontrivial take-home examinations.

A 90-item open-ended questionnaire was given to the students in the discovery class, with the following general results. On the negative side, the students did not find certain mathematical topics to be either interesting or useful. Most students were concerned for varying periods of time about covering enough material. Group conflict or frustration sometimes occurred, especially when the mathematical problems were too hard. The formation of effective and satisfying working groups was rather difficult for the students.

On the positive side, the pilot class had either positive or nonnegative effects upon each student's interest in mathematics and estimate of his or her problem-solving skill. Most students believed that working with others had positive effects on their own learning and that working problems every day did not become routine or monotonous. Almost all of the students had a closer, more personal relationship with their mathematics teacher than with their other teachers, and half of the class members saw other class members socially. Most of the students found their calculus class more stimulating than their other classes, and the attitudes of all the students toward the class either stayed the same or improved during the year.

A number of students asked for an extension of the class for another semester. The Department of Mathematics consented to schedule a third semester continuation for the small group discovery class.

On the basis of the evidence, it seems fair to conclude that the pilot study was a successful first attempt to implement the small group discovery method. A number of questions will now be presented for future investigation.

Questions for Further Investigation

This chapter describes the development and initial tryout of a new mathematics instructional system designed to foster active learning, thinking, student pacing, and interpersonal communication. A number of the questions that arose as a result of the pilot study fall into three broad areas. The three areas are concerned basically with mathematics questions, further development of the small group discovery method, and the applicability of the method to different student populations. These questions are listed here.

- Does the small group discovery method increase student interest in mathematics?

- Does the small group discovery method increase student skill in solving mathematical problems? Is that increase greater when the teacher emphasizes the use of heuristic techniques? [Editor's note: A recent study by Loomer (1976) does not support an affirmative answer to these questions.]

- Can mathematical creativity be fostered through participation, over an extended period of time, in a small group discovery approach?

- Is it possible to develop written curriculum materials in calculus for small group instruction so that students will perceive most topics as being interesting or useful? Or, in any method of instruction, will calculus students find some topics (e.g., proofs using the definition of the limit and the development of the natural logarithm as in integral) uninteresting, not useful, and difficult?

- Can calculus students improve and feel comfortable with continuous functions only if they have extensive prior experience with noncontinuous ones?

- Can a mathematical topic be understood by average mathematics students using expository instruction and by high-achieving mathematics students using a small group discovery approach over equal periods of time?

- If a mathematical topic cannot be developed by high-achieving mathematics students using a small group discovery approach, can that topic be understood by average students using an expository approach?

- What are the differences in understanding and the time required to develop that understanding between equivalent groups of students who are instructed by an expository approach and by the small group discovery method?

- Can a small group discovery method be used in mathematics instruction at all school and collegiate levels? If so, what would the effects of this technique be upon mathematics learning and the quality of interpersonal relationships?

- When the small group discovery method is used with college students, what is the optimal length for class meetings?

- What is the opitmal size of a work group in a mathematics class? Does the optimal size of a work group vary with the type of thinking or skills required to solve a problem?

- What is the optimal size of a mathematics class taught by the small group discovery method?

- What procedures can be developed to facilitate the formation of effective and satisfying work groups?

- What are the ways to improve group functioning and interpersonal relationships in a small group discovery method mathematics class?

- What grading systems are especially well suited to the small group discovery method?

- What types of students are best suited for learning using the small group discovery method?

- Can the small group discovery method function positively for students who wish to change their interpersonal style of behavior (e.g., their ability to cooperate or to share responsibility)?

- How can the small group discovery method be varied so that each group works at its own pace with a set of written materials?

Further Work

Since the pilot study, the author has used the small group method to teach courses in calculus, honors calculus, abstract algebra, transformation geometry, non-Euclidean geometry, and mathematics for elementary school teachers. In addition, some colleagues have used small groups in teaching precalculus mathematics, linear algebra, advanced calculus, complex variables, and algebraic topology. There were no special admissions requirements for these classes; any student who had the prerequisites was admitted.

After the pilot study, the author made several changes in his teaching of small groups. Class sizes were larger, typically ranging from 20 to 28 students. For classes with more than about 28 students, an assistant was needed to help supervise the groups.

The teacher frequently introduced new concepts and problems in written form, rather than in class discussions. The use of dittoed worksheets or a special text allowed each group to set its own pace when working through the materials. The teacher did not provide any notes containing a record of the students' accomplishments. Students took their own notes if they wished to do so.

Considerable care was taken in forming the work groups. At the beginning of each semester, students were asked to switch groups frequently in order to meet many different class members. After this initial period of acquaintance, students were asked to write down privately the names of class members they preferred to work with, those they could tolerate if necessary, and those they wished to avoid. The teacher then used this written information to form compatible groups which remained together throughout the semester.

The course grade was not always based on the A-B scale. The grade was based on take-home exams, attendance, and homework. The teacher checked some homework problems, and class members took turns checking others. A student could turn in the same homework problem several times until it was finally correct.

Since 1972 the author has offered an annual graduate seminar for teachers who wished to learn about using small group instruction. Participants have been mathematics teachers at the elementary, secondary, and college levels. In the seminar, the theory, practical techniques, resource materials, and research literature for small group learning of mathematics have been taught. Concurrent with the seminar, each participant has taught a course employing the small group method. The seminar has been used as a support group for teachers to try out new ideas and to resolve problems in their teaching with small groups.

Other individuals have employed the small group discovery method with differing student populations and for various reasons. Using this method, a model for developing curriculum materials has been evolved by observing the work groups (Davidson, McKeen, & Eisenberg, 1973; Eisenberg, 1970; McKeen, 1970) and learning hierarchies have been developed by small groups (Seidl, 1971; Shriner, 1970). Buchoff (1970) has reported the development of programmed materials for use with pairs of high school plane geometry students, and Jordy (1976) reported the development of small group discovery lessons for use in the Secondary School Mathematics Curriculum Improvement Study Materials (Fehr, Fey, & Hill, 1972a, 1972b). Poppendieck (1971) and Thoyre (1970) have used small group methods in teacher education courses. Several studies (Brechting & Hirsch, 1977; Davidson & Urion, 1977; Gallicchio, 1976; Grant, 1975; Hildenbrand, 1975; Kenney, 1974; Klingbeil, 1974; Klingbeil & Davidson, in press; Loomer, 1976) have attempted to determine the effects of using the small group discovery method.

Research on the small group discovery method has been supplemented by developing text materials especially designed for use with that method. Thus far materials have been developed for courses in elementary algebra (Stein & Crabill, 1972), plane geometry (Chakerian, Crabill, & Stein, 1972), abstract algebra (Davidson & Gulick 1976), mathematics for elementary teachers (University of Maryland Mathematics Project, 1978; Weissglass, in press) and mathematics for liberal arts majors or elementary teachers (Knaup, Smith, Shoecraft, & Warkentin, 1977). At present, text materials are being prepared for courses in linear algebra (Dancis, unpublished manuscript) and the calculus of one variable (Leach & Davidson, unpublished manuscript).

A more complete description of the small group discovery method and of the pilot study can be found in Davidson (1971a). Other published papers dealing with the small group discovery method are Davidson (1971b), David-

son (1974), Davidson (1976), Davidson, Agreen, and Davis (1978), McKeen and Davidson (1975), and Weissglass (1976, 1977).

58

Chapter 4

Development of a Unit of Number Theory for Use in High School, Based on a Heuristic Approach

Shlomo Libeskind

The Problem and Its Background

"How does it happen that so many refuse to understand mathematics?" Poincaré asked (1929, p. 43; 1969a; 1969b, p. 295) at the beginning of the century. This question is at least as relevant today as it was then.

In spite of recent efforts to develop new curriculums, textbooks, and materials, the number of students failing or doing badly in mathematics is enormous. It is common to hear students at all levels, high school and college, complain that mathematics is a dry uninspiring subject and that it depends upon many incomprehensible tricks.

Proof and deductive reasoning are at the very heart of mathematics, yet in textbooks and classrooms, mathematics is usually presented as a finished product. The student is rarely told how one starts a proof or proceeds from one step to the next. As a result many find it difficult to reproduce proofs they have learned and almost impossible to prove new statements and solve more challenging problems.

Traditionally the student's first encounter with proof is in high school geometry (usually tenth grade). Some newer curriculum programs present proof along with algebra, although proofs in beginning high school algebra involve field axioms and are difficult for most students, even when well presented. Later in algebra, emphasis is placed on techniques for solving particular types of problems, and even when proofs are presented they are rarely emphasized. The problems that most students are able to solve are usually routine. In geometry where students encounter proofs and more challenging problems, they also experience more difficulty.

In view of this situation this author wanted to develop a unit on proof for use in high school that would be accessible to students with a background in beginning algebra. Thus it was decided to develop a unit in number theory using a specially designed heuristic approach, based on the teaching and learning of problem solving advocated by Polya (1954a, 1954b, 1962, 1965). When using this approach in proving a theorem or solving a problem, a teacher does not merely justify each step by referring to a previously proved theorem, defi-

59

nition, or axiom but shows why it is reasonable to start the proof in one way and not another and how one knows how to proceed from one step to the next.

The overall objectives of the study were to develop such a unit in number theory and to test its feasibility by presenting it to an ungraded class of high school students. Three basic questions were asked: (a) Can the students reproduce the proofs of the theorems in the unit? (b) Can the students understand the meaning of the theorems? (c) Can the students apply the methods used in proving the theorems in the unit to solve new problems which include proving statements the students have not seen before?

Development of the Unit

The development and tryout of the number theory unit were based on a curriculum development model advocated by Romberg and DeVault (1967). According to that model, the steps in developing an instructional system are analysis (mathematical and instructional analysis), pilot examination, validation, and development. The study carried the development of the unit only through the first two phases of this model. As Romberg and DeVault point out these two phases are of great importance in the development of an instructional system (1967, p. 107).

Mathematical and Instructional Analysis

In order to keep the mathematical prerequisites for the unit minimal, it was decided to work within the system of whole numbers. Thus, the symbol $d|a$ was defined as follows: $d|a$ if there is a whole number k such that $a = kd$. The following main theorems and topics were chosen:

Theorem 1: If $d|a$ and $d|b$ then $d|(a + b)$.

Theorem 2: If $d|a$, $d|b$ and $d|c$ then $d|(a + b + c)$.

Theorem 3: If $a > b$, $d|a$ and $d|b$ then $d|(a - b)$.

Theorem 4: If $d|a$ and k is a whole number then $d|ka$.

Theorem 5: If $d|a$ and $d|b$ then $d|(ka + nb)$ where k and n are whole numbers.

Theorem 6: If $a|b$ and $b|c$ then $a|c$.

Divisibility by 2, 4, and 5.

The meaning of "if and only if."

Theorem 7: If $d|a$, $d\not|b$ then $d\not|(a+b)$.

Theorem 8: There are infinitely many primes.

Theorem 9: If n has no prime factors less than or equal to the square root of n, then n is prime.

Sieve of Eratosthenes.

Theorem 10: $(a, b) = (a - b, b)$ where (a, b) denotes the greatest common divisor of a and b.

Euclidean algorithm.

A task analytic approach developed by Gagné (1965) guided the development of the unit. This approach is well described by King (1970):

> The idea is to express the objectives of instruction in terms of observable performance tasks. If the instruction is successful, the students will demonstrate the ability to perform the specified behavioral objectives. Hence, the success of the instruction is measured in terms of student performance on predetermined performance objectives. Once the curriculum developer has specified these objectives, a task analysis is performed. The task analytic procedure is performed. The task analytic procedure was developed by Gagné to train human beings to perform complex tasks. The basic idea of this approach is to break down each behavioral objective into prerequisite subtasks; these subtasks may in turn be analyzed into finer subtasks. The procedure continues until one reaches a set of elemental tasks which cannot or need not be further analyzed. If properly done, the task analysis should yield a hierarchy of tasks which indicate the steps a student must take in order to learn the terminal behavioral objective. The hierarchy indicates how instruction would proceed: one starts with the simplest tasks and learns each subtask until the terminal objective has been mastered. (pp. 48-49)

Any proof or solution to a problem has two basic components: (a) knowledge of an ability to manipulate subject matter content, and (b) a plan or strategy which permits the student to use the subject matter content to form a valid argument.

The task analysis related to the first component is usually quite simple to identify. The proof of Theorem 1, for instance, needs the application of the definition of *divides*, the substitution principle, and the distributive law.

The second component, the plan or strategy, is of utmost importance. Being able to find a strategy makes the difference between finding a proof (or solving a problem) or not finding one. Here are the greatest difficulties that most students encounter. As already pointed out, some textbooks outline a plan or strategy for proving a theorem or solving a problem, but very often these plans are merely recipes for solutions and do not explain to students why this plan was chosen and not another, or why each step within the plan was taken. In this way most plans fail to show students how they should go about finding a proof or solution on their own.

The present study put great emphasis on showing how strategies could be found. Thus a modification of Polya's (1945) heuristic approach was used. A similar approach was used by the present writer in two expository works: Libeskind (1968) and Beck, Bleicher, and Crowe (1970, in particular Chapter 2). The main idea in the approach was to show the student why it is reasonable to take each step and to point out alternative approaches.

One of the goals of the study was to instruct students in the use of the heuristic process, that is, to encourage them to ask heuristic questions when confronted with reproducing proofs of the theorems and solving new problems. To achieve this, it was decided to encourage active student guessing while proving the theorems and solving problems. Students were asked to suggest what the next step in a particular proof should be. To avoid situations in which one student responds to a question and the teacher continues as if the whole class responded, the response of more than half of the students was sought by asking the students to write answers in their notebooks. Students who did not get the answer were given further hints.

To discourage memorization, proofs were written in several different forms: two column, story type, a combination of these two types, and diagram form. Writing proofs in different forms is also valuable for other reasons. A two-column proof is helpful for beginners, since it is structured in the form Statement—Reason, and reminds the student to give a corresponding reason to the statement. A story-type proof is universally used in mathematics as it is an easy way to write and explain a longer proof. A diagram approach is sometimes useful in discovering a proof.

Often the same theorem or problem was proved or solve by several different methods. The reasons for this are:

- A student who does not see one approach may find another understandable.

- One important general principle appears to be this: wherever possible, the child should have some intrinsic criterion for deciding the correctness of answers, without requiring recourse to authority. . .

- In more advanced work in later grades, solving problems by several different methods, recognition of patterns, and even the use of simple logic will play the role of foundation for deciding correctness without recourse to authority (Cambridge Conference on School Mathematics, 1963, pp. 15-35).

The second goal of the experiment was to demonstrate that students could master the objectives of the unit. To accomplish that goal, the idea of *mastery learning* was used. This assumes that given enough time all or almost all students can learn the intended course material and it is the task of the instructor to find the means and methods to obtain this mastery. Mastery learning has been extensively reviewed in a book edited by Block (1971). King (1970) and Shepler (1969) demonstrated that mastery learning in mathematics can be successfully used in elementary schools.

For this study, the mastery criteria were: (a) a student must respond correctly to at least 75% of the items on the test in order to be considered a master, and (b) at least 70% of the students had to be considered as masters, on all tests.

In order to achieve mastery learning of the behavioral objectives of this instructional unit, the following were used in the studies:

1. A heuristic approach was employed.

2. On each test students were rated as masters or nonmasters. If a student was a nonmaster on a topic, that student was given further instruction and an opportunity to take a parallel test. If graded a master on a second or third test he or she was counted as a master for the topic.

3. Seven booklets were developed which the students used to learn and review most of the theorems in the unit. On each page of the booklets there was expository text and questions. Students were asked to answer the questions and compare their responss with the answers on the following page of the booklet. Sometimes immediately after the question there appeared the word "Hint" with a number after it. In that case the students could use the corresponding hint on the last page of the booklet, but only after trying unsuccessfully to answer that question.

4. Problem sheets were given daily as homework; these were corrected and mistakes were pointed out. Solutions and mistakes were discussed in class and individually if necessary.

5. If the students were masters on all the mastery tests they would be considered masters of the unit.

The Pilot Study

To aid the development of materials appropriate for high school students, a pilot study was conducted in the summer 1970. The pilot consisted of 25 sessions, including testing sessions, of about 50 minutes each. After each session there was a study period of about 40 minutes. Five black students, two girls and three boys, from inner-city schools in Michigan were taught by the author. These students were participants in the Michigan State University Inner City Project (MSUIC-MP). They were average and above average students; two had finished ninth grade, one had finished tenth, and two the eleventh grade. The students were selected on their teachers' recommendations that they were probably of college capability.

A test of prerequisites was administered. It showed that most of the students, especially post ninth- and tenth-grade students, needed instruction in the prerequisites. So, a 1-week unit on the prerequisites was given before teaching the 4-week experimental unit itself.

The results on the problem sheets and the mastery tests in the pilot study demonstrated that the unit was appropriate for an average or above average ungraded group of high school students, but the experience gained in teaching the unit showed some minor changes were desirable. Some problems

were added in the problem sheets and a few explanations in the booklets were clarified. A seventh booklet with a different proof of Theorem 10 was added. In proving Theorem 10, two students had suggested considering the set of all common divisors of $a - b$ and b, and showing that the sets are equal. Along these lines booklet No. 7 was developed.

Another change was made in the order of presentation of Theorem 9 and the Sieve of Eratosthenes. In the pilot study the Sieve method was presented first, in the hope that the content of Theorem 9 would be discovered from the Sieve method. However, students had difficulty discovering Theorem 9 even after hints were given. It was decided to try another method in which Theorem 9 was done first. The second method worked much better, so it was used in the main study.

The Main Study

As pointed out, this study was based on a curriculum development model developed by Romberg and DeVault (1967). In the main study, conducted during the summer of 1970, 10 average and above average students, seven girls and three boys, were taught the experimental unit by the author. As in the pilot study, the students were participants in the Michigan State University Inner City Mathematics Project (MSUIC-MP) on the campus of Michigan State University. Nine were black, and one was white. Three students had finished the ninth grade, three the tenth, and four the eleventh grade. The students were selected for the summer institute on their teachers' recommendations that they were probably of college capability.

The main study consisted of 25 sessions of about 50 minutes each. After each session there was a study period of 30 minutes. On the first day students were given a test on prerequisites and a pretest. As in the pilot study, the test results showed that it was necessary to spend the first week teaching prerequisites. Two experienced high school teachers were present at each session. They observed and wrote a protocol of all activities and their notes were used in writing journals for the lessons (Libeskind, 1971, pp. 21-224).

Conclusions

Nine students took the posttest in the main study; each of these students were masters on the posttest as well as on the four mastery lesson tests. The test results were used to answer the three basic questions posed earlier:

1. Can the students reproduce the proofs of the theorems? The results of the mastery tests and posttest showed that the students were able to reproduce the proofs.

2. Can the students understand the meaning of the theorems? The results of the tests indicated that the students understood the meaning of the theorems. The students were able to give numerical examples of the theorems and apply them to divisibility facts. They were able to use some of the theorems to use certain algorithms (finding if a number is prime, Sieve of Eratosthenes, or the Euclidean algorithm).

3. Can the students apply the methods used in proving the theorems in the unit to solve new problems which include proving statements the students have not seen before? The results showed that the answer to this question is in the affirmative as well.

In regard to Question 1, the data shows that the students were able to reproduce the proofs even though they were not asked to memorize them. In fact, the students were explicitly discouraged from memorizing the proofs. Thus, the ability of the students to reproduce the proofs may owe much to the method of instruction and the use of the heuristic approach.

The affirmative answer given to Question 3 is of particular significance. The ability to solve new problems and prove new statements was considered a transfer measure of the understanding of the proofs in the unit and a measure of the success of the heuristic method of instruction. A part of the pretest-posttest was primarily concerned with this type of new problems.

The ability of the students to recognize if a proof is valid was another indication that the students understood the proofs in the unit and did not just memorize the steps and their reasons. These results are especially encouraging since proofs of theorems such as Theorem 9 and 10, the Euclidean algorithm and some of the problems in the problem sheets, and mastery tests are difficult even for college students.

The seven booklets played an important role in learning the theorems in the unit. Usually after completing a booklet, the students were confident that they knew the proof of the theorems in that booklet.

The students enjoyed the unit. Most were active in classroom discussions. They particularly enjoyed the application of Theorem 9 to the search for prime numbers, the Sieve of Eratosthenes, and the application of Theorem 10 to the Euclidean algorithm.

The students reacted positively to the idea of mastery learning; many remarked that they would like this procedure used in their schools. All the students were eager to become masters the first time they took a test. The mastery learning procedure worked well in a small class situation, where the teacher could see the progress of each student and give individual help when necessary.

Recommendations for Further Study

The heuristic approach used in the unit seems very promising, although the study carried the development of the unit only into the Pilot Examination phase of the developmental model of Romberg and DeVault (1967). However, at the beginning a strong Hawthorne effect was evident. This effect seemed to be due to the students' new university environment and awareness that they were in a special project, rather than to the experimental nature of the unit.

Thus, the findings must be subjected to further examination. The results suggest the following recommendations for further study:

1. The development of the materials in the unit should be continued to determine whether the unit will be effective with other groups of average and above average students, and whether other teachers will be able to teach it using the heuristic approach. The necessity for this is aptly pointed out by Romberg and DeVault (1967):

> Assuming that a procedure has proven to be feasible in its pilot-tryout, the next phase is validation. The materials and methods need to be tried out in a variety of regular classrooms with other kinds of learners, other kinds of teachers and in different social contexts. (p. 108)

2. The success of this study suggests that it might be feasible to design and teach other material this way. Since the students particularly enjoyed the applications of Theorems 9 and 10, the extended unit should include congruences, divisibility tests by 3, 9, and 11, Fermat's theorem, and some number theoretic functions.

3. It would be valuable to find out if the heuristic approach used in the study could also be used in the elementary school. King (1970) designed a unit on proof for sixth grade and showed that the students were able to apply the theorems in his unit to simple divisibility facts and to reproduce the proofs of these theorems. The students in King's study were drilled on the proofs of the theorems. They were able to reproduce the proofs, but they were only asked to prove three statements they had not seen before. It would be interesting to find out whether the heuristic approach used in the present unit would enable sixth-grade students to reproduce the theorems with less drill, and result in transfer and successful problem solving on a higher level.

4. The effectiveness of the heuristic approach suggests that developing materials using such an approach for high school and college classes could be worthwhile.

Chapter 5

An Exploratory Study on the Diagnostic Teaching of Heuristic Problem-solving Strategies in Calculus

John F. Lucas

In a 1976 paper on basic mathematical skills prepared for the National Institute of Education by the National Council of Supervisors of Mathematics (1976), the following statement is especially noteworthy: "The main reason for studying mathematics is to learn to solve problems."

The study outlined in this chapter was conducted 6 years earlier (Lucas, 1972), but it was motivated by precisely the same assumption. If the assumption is true, researchers in mathematics education and mathematics teachers ought to be searching for ways to improve learning, teaching, and communicating mathematical problem solving. This can be accomplished most effectively through collaborative efforts of researchers and teachers. The major difficulty is where to start looking.

In conceiving this study, the writer focused on what he saw to be the psychological core of mathematical problem-solving, i.e., heuristics. Heuristics are higher-order, tentative, general decision processes which help organize and narrow the search for a problem solution. Drawing diagrams, separating information, reasoning backwards, recognizing and using analogies, searching for patterns, successive approximation, checking, and exploiting problem symmetry are some examples of heuristic behavior. These actions are tentative in that they do not guarantee success; they are general in that they apply across specific problems and classes. In contrast, processes such as applying the quadratic formula to solve certain equations or Gauss-Jordan elimination to reduce a matrix are algorithmic, since they are always successful when correctly applied and since they apply only to specific kinds of problems. Heuristics, on the other hand, transcend classes of problems and are a unifying element in the study of problem-solving. Information about teaching and learning heuristics is critical for understanding the entire process of solving mathematical problems and its relationship to teaching and learning mathematics. This assumption played a significant role in the development of this study.

At the time of the study reported here, few research studies had been concerned specifically with heuristics. The mathematician George Polya (1948, 1954a, 1954b, 1962, 1965) provided a great quantity of information on heuristics in many interesting mathematical problems and discussions. Polya has furnished mathematics educators with material to be tested in many dissertation studies, including this one. Essentially he condensed his own exper-

iences and those of other writers, distilled the key ingredients of mathematically oriented mental processes in problem-solving, and arrived at an array of heuristic processes. Polya's writings demonstrate the utility and effectiveness of heuristics in the hands of skilled problem solvers. His observations led to an inquiry-oriented model for teaching mathematics, where the teacher asks heuristic questions and makes suggestions, and the learner develops self-direction by asking the same questions while attempting to solve problems.

Although the heuristics identified by Polya are important objects for research, it is difficult to gather evidence about them since they must first become observable actions, and there must be a system for recording their occurrence and making measurements. Traditional paper-and-pencil tests are clearly inadequate for observing the problem-solving process or measuring its content. However, the problem of direct observation is largely overcome by requiring the problem solver to think aloud as he or she works. This technique was used by Duncker (1945), by information-processing theorists (see Newell, Shaw, & Simon, 1958), and more recently by Soviet investigators (see Krutetskii, 1969).

In the late sixties, Jeremy Kilpatrick (1968) developed a system for coding problem-solving actions and events observed from tape-recorded thinking-aloud protocols. In Kilpatrick's system, symbols representing behaviors and events were recorded in the same sequence as those events actually had occurred. Kilpatrick called this a process-sequence code. Using this system, an observer could record a time-exposure snapshot of problem-solving actions. In the study outlined here, Kilpatrick's system was considerably modified with respect to content, but its structure nevertheless provided the basic idea for gathering data.

Lacking a firm theoretical base on which to build, the study reported here was planned as an exploratory study. It was partly influenced by the "teaching experiment" style of Soviet problem-solving research. Research on problem-solving processes in mathematics was in its infancy, with a great deal of exploratory work, observation, treatment variation, data probing, and conjecture needed before rigorous experimentation could be executed (Kilpatrick, 1970). Thus in 1970, exploratory work aimed at the mundane task of developing better behavioral analysis instruments would be more beneficial than ambitious attempts to define effective teaching or learning of problem-solving. Thus the major objective of this study was simply to develop conjectures to help set a course for further investigation.

The study took place in a first semester calculus course (differential and integral calculus of real-valued functions of one variable). This setting was chosen because student subjects were accessible to the investigator, and because the calculus offers a challenge in terms of integrating problem-solving techniques with standard content. The plan was to conduct a clinical diagnostic teaching experiment in which as many of Polya's heuristics as feasible

would be introduced to freshman calculus students. The practice (treatment) problems were calculus problems; the experimental (test) problems were general nonroutine mathematical problems. Observation was conducted in audio tape-recorded individual sessions in which subjects thought aloud while solving problems. Protocols were analyzed through a system modified from Kilpatrick's and developed specially for this study. The data were probed; outcomes were conjectures pertaining to the following researchable questions:

1. Is it possible to teach heuristics and produce observable effects? If so, what are the nature of the effects?

   a. Are there strategy shifts? Is there an increased frequency or a change in emphasis on heuristics?

   b. Are there changes in problem-solving performance? Are there effects on time, accuracy, completeness, difficulty, or errors?

2. Can Kilpatrick's system of behavior analysis be adapted for observation of college students?

3. Is it possible to devise a reliable modification of Kilpatrick's system?

4. Can heuristics be integrated into college calculus without sacrificing course content?

The author believed that answers to these questions would help guide researchers, generate ideas for classroom teachers, and improve communication of mathematical problem solving. It is the purpose of this paper to describe a heuristic teaching experiment conducted by the author in 1970, a method for analyzing problem-solving processes, a summary of tentative conclusions, and further work in research and curriculum development undertaken by the author as a consequence of this study. It is not the purpose of this paper to emphasize inferences that might be drawn from the data. The reader interested in greater detail is directed to the corresponding dissertation or journal report (Lucas, 1972, 1974).

The Study: Structure and Design

The main study was preceded by a pilot study in spring 1970 with six volunteer students who were taking a second semester calculus course from the author. These students were given a set of 15 mathematical problems, mostly rate and optimization problems in differential calculus. They were asked to solve the problems while thinking aloud. Several interview sessions were required in all but one case, and the total recorded time averaged 4.25 hours per subject. The purposes of the pilot study were to familiarize the investigator with the interview procedure, provide process-sequence samples for coding practice and revision, and determine which heuristics were most likely to be observed in student problem-solving at that level. As a consequence of the pilot

Table 1

Design of the Study

Group	*N*	I (Pretest)	II (Instruction)	III (Posttest)
H_1	8	O[a]	H[b]	O
C_1	6	O	no-H	O
H_2	9	no-O	H	O
C_2	7	no-O	no-H	O

[a]O = diagnostic observation.

[b]H = instruction on heuristics.

study, the coding system and interview format were revised several times in preparation for the main study.

In fall 1970, 30 university students from two first-semester calculus classes taught by the investigator participated as unpaid volunteers in the study. The study was executed in three phases during the 14-week term. Phases I and III were 2-hour diagnostic observation interview sessions (testing) with individual subjects. This series of interviews (pre- and posttests) lasted about 2 weeks each. Phase II was an 8-week instructional treatment. A Solomon four-group design (Campbell & Stanley, 1969) involving two treatment conditions (explicit heuristic instruction × no explicit heuristic instruction) and two testing conditions (pre- and posttests × posttest only) was used. Table 1 illustrates this design.

One class was taught using explicit emphasis on heuristic techniques, and the other served as control with no explicit reference to heuristics. Subjects were not randomly assigned to treatments, but appeared in a treatment group by registering for that section of the course with no foreknowledge of an experiment. The two groups were taught one after the other each morning. The total number of subjects (30) was small because of the individualized nature of the interviewing sessions and the amount of time required for analysis and coding of observations from 44 2-hour sessions. Background information, including age, sex, class, intended major, semesters of high school mathematics, grade point average, and ACT mathematics percentile rank, was obtained for each subject. On these particular traits, all four groups were very similar.

During Phase I, 14 subjects participated in problem-solving interview sessions. Each subject was given a booklet containing instructions, two sample problems, and seven test problems. These problems were general (noncalculus, except for the last two), and each had several potential solutions. Two examples are given.

Problem 1 (Pretest) The speed of sound in an iron rod is 16,850 ft/sec, and the speed of sound in air is 1,100 ft/sec. If a sound originating at one end of the rod is heard 1 second sooner through the rod than through the air, how long is the rod?

Problem 2 (Pretest) A real estate agency offers you a choice of two triangular pieces of land. One piece has dimensions 25, 30, and 40 feet; the other has dimensions 75, 90, and 120 feet. The price of the larger piece is 5 times the price of the small piece. Which is the better buy?

The subject was told to solve the problems using pencil and paper, but to think aloud while working. These remarks were tape-recorded, and the interviewer noted various observations. Interaction between interviewer and subject was minimal, except for an occasional reminder to think aloud whenever the subject lapsed into silence. Retrospective comments by the subject or feedback from the interviewer about correctness and quality of the solutions were avoided. For each interview, the record of thinking aloud, the subject's written work, and the interviewer's notes formed a collage of information representing the problem-solving process. This information was studied and reduced to a checklist, process-sequence code, and score, which are described next.

Behavioral Analysis: A Coding System

Integrating the structure of Kilpatrick's behavioral analysis (1968), a model of heuristics for solving mathematical problems, and the experience and observations of the pilot study, the investigator created an extension of Kilpatrick's system which included those of Polya's heuristics observed during the pilot study. This new system consisted of a checklist (see Figure 1), a process-sequence code (see Table 2), and provisions for scoring various aspects of time consumption and general performance. The checklist categories included several heuristics not represented in the process-sequence code, but the major function of the checklist was to provide a more detailed analysis of some heuristics and events which were assigned process-sequence codes. In addition, three measures of time and four measures of score were taken for each problem.

Of particular interest in the study were heuristic strategy shifts, changes in nature or frequency of heuristics, and changes in problem-solving performance. To detect these changes, a system of behavioral analysis was designed to record and evaluate many actions which could occur during a problem solution. The kind of notation used, the number of diagrams drawn, whether or not the diagrams accurately represented problem conditions, the number and kinds of diagram modifications, and whether or not the subject recalled a related problem or applied its method or result were examples of

CODING FORM (FINAL VERSION)

Subject No. _____ Time: exc. looking back _____
Problem No. _____ looking back _____
Coder _____ Tape No. _____ total _____
Date _____ Tape Readings _____ Score: approach _____
plan _____
result _____
total _____

Approach
- restates problem in own words _____
- mnemonic notation _____
- representative diagram yes _____ no _____
- auxiliary line (s) _____
- enlarges focal points _____

Production
- recalls related problem _____
- uses *method* of related problem _____
- used *result* of related problem _____
- inductive reasoning _____
- (pattern search) _____

Looking Back
- routine check of manipulations _____
- is result reasonable? _____
- all information used? _____

Checking
- test for symmetry _____
- test of dimensions _____
- specialization (extreme cases) _____
- comparison with gen. known result _____

V_s
- condenses/outlines process _____
- tries to derive differently _____

V_m
- variation by analogy _____
- variation by changing conditions _____

Executive errors (tally)
- algebraic manipulation _____
- numerical computation _____
- differentiation _____
- other _____

Interviewer Comments

PROCESS SEQUENCE

Figure 1. Checklist categories coding form.

some of these activities. The frequency of checking was measured by a process-sequence code; the kind of checking was classified by seven checklist categories. Similarly, two process codes were used to distinguish and sort errors of structure and execution; the checklist further distinguished four categories of executive errors. Instances in which errors were noticed and corrected were also counted. Strategies by which a solution is produced (e.g., analysis, synthesis, trial and error, reasoning by analogy) had corresponding process-sequence codes. The by-products of the solution (e.g., equations, relations, and algorithmic processes) were also recorded. Another process code was used if the subject was observed separating or summarizing problem data. The system also had codes for looking-back behaviors such as condensing or outlining

Table 2

Process-sequence Codes

Code symbol	Observed behavior
R	*Reads* Problem
S	*Separates/summarizes* data (wanted vs. given; relevant vs. irrelevant)
M_f	Model introduced via *figure*, diagram, schematic
$M_{f'}$	*Modification* of existing figure (auxiliary lines; enlargement; darkening, etc.)
M_{f_c}	Model introduced via figure with *coordinate system*
DS	*Deduction* by *synthesis* (working forward)
DA	*Deduction* by *analysis* (working backward)
T	*Trial* and error (successive approximation)
An	Reasons by *analogy* (using methods, results, ideas from problems similar in structure)
Me	Model introduced via *equation*(s) or other algebraic relationship
Alg	*Algorithmic process*
N	*Nonclassifiable* behavior (mumbling, incomplete statements, random guessing, etc.)
C	*Checks* result
V_s	*Varies* the *process* (attempts alternate attack)
V_m	*Varies* the *problem* (invents new related problem)
↯	*Structural error* (misinterpretation; misrepresentation)
↓	*Executive error* (manipulative error; miscalculation)
•	*Error* explicitly *corrected*
-	*Hesitation* of two units (30 seconds)
/	*Stops without solution*

a solution process, trying a different mode of attack, or inventing a new problem related to the given one. Still other code symbols indicated difficulty, namely, hesitating, rereading the problem, and stopping short of a solution. When the composite picture was reconstructed from a tape-recorded vocalized protocol was examined very carefully, little observable behavior was likely to escape scrutiny. There were 25 checklist categories in the original system and 20 process-sequence codes. The latter are presented in Table 2.

In applying the codes listed in Table 2, parentheses were used to cluster subprocesses related to a more general process, commas separated processes, and a period denoted a completed solution. Outcomes of processes were indicated by numerals 1 through 5, respectively, for abandons process, impasse, incorrect final result, correct final result, and intermediate result. To illustrate its appearance, the coding string

R, M_f, -, DS (Me,Alg)5, DA (Alg)5, C, DS (Me,Alg)4,C.

would be translated as follows: The subject read the problem (R), drew a figure (M_f), hesitated at least 30 seconds (-), started putting information together (DS) to yield an equation (Me) which was solved by a standard technique (Alg) to obtain an intermediate result (5). Next, the subject looked at the goal and asked what was needed to obtain that (DA), followed by a brief calculation (Alg) in which a mechanical error (•) was made. Upon checking back (C), the error was discovered and corrected (*), and the subject proceeded in a forward manner (DS) to derive another equation (Me) which was solved (Alg) to produce a correct final solution (4). This solution was verified by checking (C) against the conditions of the problem.

Using the system demonstrated above, in combination with a checklist for further clarification, the investigator was able to record the evolution of the problem solution so that a much clearer picture could be obtained than that afforded by written work alone.

Evaluation of problem-solving performance included measures of time, score, difficulty, and errors. Time was measured in unit intervals of 15 seconds each, and three time measures were taken: time excluding looking back (a performance measure), time looking back (a heuristic measure), and total time (sum of the two). The solution score was split into four weighted parts, corresponding to approach (subject demonstrated understanding of problem, 1 point), plan (subject derives information sufficient to solve problem correctly in the absence of executive errors, 2 points), result (subject establishes correct and complete result, 2 points), and total score (sum of approach, plan, and result scores, 0-5 points). Difficulty was inferred from the frequency of hesitation, rereading of the problem, impasses, and stopping without the solution. Finally, errors were subdivided into two types, structural and executive, and the latter were further subdivided in the checklist. These measures helped produce a composite picture of the problem solver's general performance.

The system of behavioral análysis described above was applied to the observational data from all interview sessions of Phases I and III, having 14 and 30 sessions, respectively. This analysis required approximately 400 hours of work and was carried out in the semester following the completion of the study.

The Instructional Program

Phase II, the instructional program and the heart of the study, spanned a period of 8 weeks, or forty 50-minute class periods from September 30, 1970 to November 25, 1970. During this phase, calculus topics in both classes included limits, continuity, the derivative, differentiation, applied problems involving derivatives, and the antiderivative. These concepts and related mathematical information were introduced in the same expository manner in both

classes. The differences between modes of classroom teaching centered on the nature of problem assignments, the depth of discussion of problem solutions, the grading of problem solutions, and the explicitness of reference to heuristics.

All problem assignments for the control class were made from the textbook (Leithold, 1968). Since answers to problems were available to students, the control class assignments were not graded. However, a part of each class period was set aside for answering students' questions about problems.

In contrast, the experimental class was assigned drill exercises from the textbook and supplementary sets of calculus problems not included in the text. These supplementary sets, averaging five problems each, were prepared and sequenced in advance of the course to highlight and reinforce certain heuristics and correspond to classroom topics. Homework problems from the supplementary assignments were graded several times each week, and the instructor's written comments included heuristic suggestions. The grading system itself was designed to reward use of heuristics. For example, outlining key points of a solution, producing alternate solutions, or posing related problems received extra credit.

During the instructional period, the control class was assigned about 20% more problems than the experimental class but problem solutions were discussed differently in each class. The instructor responded to students' questions in the control class, whereas in the experimental class he guided their questions by raising issues and making suggestions intended to draw attention to heuristics.

Explicit introduction of heuristics was avoided in the control class and emphasized in the experimental class. A set of 12 "Heuristic Papers" was prepared by the author in advance of the study. These were distributed to the experimental class as additional reading at intervals throughout Phase II. Each heuristic paper made one or more heuristic techniques explicit through carefully constructed applications to both calculus and noncalculus problems. Historical comments and discussion of the value and effectiveness of heuristic techniques in mathematical reasoning were emphasized in each paper. A list of the titles of these papers appears in Table 3.

There was also a difference in philosophy of instruction between classes. The author believed that more effective teaching of heuristics would occur if a few problems were analyzed thoroughly than if many problems were discussed superficially. This point had been emphasized by Larsen (1960). Problem discussions in each class reflected this difference.

Another philosophical position that separated experimental from control instruction was the emphasis on asking the questions "why?" and "what if. . . ?" Buck (1965) stressed the importance of this attitude in teaching mathematics. As a consequence, students in the experimental class were en-

Table 3

Heuristic Papers

A	The Nature of Heuristics
B	Polya's Question List
1	Analysis-Synthesis
2	Method-Result Heuristic
3	Looking Back
4	Drawing Diagrams
5	Understanding the Problem
6	Checking
7	Reasoning by Analogy
8	Setting Up Equations
9	Induction
10	Sketch of a Solution Process: A Summary

couraged to be active participants in the problem-solving process rather than passive spectators, to explore and speculate rather than formalize, and to work in teams as well as independently. Voluntary 2-hour problem sessions each week, led by the investigator, were available to both the control and experimental class members.

For objectivity and diagnosis of teaching, a daily log was kept on each class. This log included notes on topics, specific examples, problems, heuristic suggestions, and questions posed by the teacher and by the students. When the student-teacher interaction in each setting was analyzed and compared, differences which were planned to be sharp became blurred; they were more a matter of degree. Problems were discussed in both groups, but the discussion provided instruction on heuristics in the experimental setting, while it reinforced concepts in the control class. Questions were asked of both groups, but they embodied general heuristic suggestions in the experimental class and were directed to specific points in the control class. Students were active participants in both groups, but experimental students were encouraged to explore, conjecture, and guess, while control students had to formulate their own questions without being asked thought-provoking questions. The teacher guided activities in both groups, but laterally in the experimental class and centrally in the control class. These were the distinctions between an instructional process which emphasized heuristics and one which did not.

Immediately after the instructional phase, a series of 2-hour interview sessions (Phase III diagnostic observations) were administered to 30 subjects, the 14 who had participated in Phase I and 16 additional volunteers from the complementary sets of students in the two classes. The two new groups exhibited similar measures on the traits used to compare the original Phase I

groups. The Phase III format was identical to Phase I except that the seven test problems were different, though several were structurally similar. Again, test problems were noncalculus oriented, except for two. Two examples taken from the posttest are given:

Problem 1 (Posttest). A man fires at a target, and 2 seconds after he fires, he hears the sound of the bullet striking the target. If the bullet travels at a speed of 1,100 ft/sec, how far away from the man is the target?

Problem 3 (Posttest). A circle whose center is at the point (4,6) is tangent to the line $3x = y - 4$. What is its area?

The posttest phase took 2 weeks. During the spring of 1971, all observational data from recorded interview sessions in both Phases I and III were assembled and analyzed by the system described earlier in this chapter. Before data analysis, modifications in the coding system had to be made as a consequence of testing for intercoder reliability, which is described next.

Reliability of the Coding System

The system of behavioral analysis which emerged from the pilot study had 52 different classifications—20 process-sequence codes, 25 checklist categories, and 7 time/score performance measures. Nineteen of these classifications represented actions or events which were clearly defined and easy to identify or evaluate. For examples, time measurements, reading the problem, drawing a diagram, and producing an equation are actions on which several coders would generally agree. On other classifications such as differences in modes of checking, the nature of errors, whether a subject was reasoning forward or backward, and assignment of a performance score, one might not expect close agreement by different judges.

To test the general reliability of the system, a faculty colleague in the Mathematics Department at the University of Wisconsin—Oshkosh was trained as an alternate coder. This training was focused on 33 potentially ambiguous classifications, under the assumption that if reasonable agreement between coders could be attained on the latter, then including the 19 clearly defined classifications would not reduce the degree of agreement. The potentially ambiguous classifications included all 25 checklist categories, the process-sequence codes DS (working forward), DA (working backward), T (trial and error), and ↓ (structural error), and four performance score measures of approach score, plan score, result score, and total score.

After a 1-week training period, 38 posttest protocols were coded by the alternate coder over a period of 5 days. These were compared with the same protocols coded by the investigator. The following three tests were applied to decide which classifications should survive for inclusion in the final system: (a)

frequency of observation; (b) coder bias (treatment effect estimates of coder differences and score); and (c) intercoder agreement.

Behaviors such as inductive reasoning, test for symmetry, and inventing new problems were dropped due to low frequency of observation. The relative absence of these behaviors in the protocols of college students made the investigator curious. Are certain heuristics problem-specific or class-specific? Perhaps the structure of the problem or the nature of the question evokes pattern-search (induction). Also, are looking-back actions (e.g., inventing new problems and trying alternate approaches) a function of student habit or does their absence relate to a situation variable like the nature or length of an interview session?

Certain classifications in the checklist, such as checking and executive error categories, were intended to refine components of the coding scheme (C, ↓). It was thought that clustering these categories together to produce a more macroscopic system would yield increased intercoder reliability. Paradoxically, this was not the case. Clustering four checking classifications having indices of agreement .81, .69, 1.00, and .69, respectively, produced a single checking variable with index .75. Similarly, clustering executive error categories with indices .84, .96, .85, and .79 led to a single variable for executive errors whose index was .83. These results indicated that a well-trained coder familiar with the mathematical setting of the problems can discriminate adequately among certain closely related behaviors.

To test intercoder agreement on process-sequence behaviors, the coders' strings of symbols were used along with coder notes on interpretation and rationale to obtain frequencies of agreement and disagreement. An agreement was tallied when the same action (as described in the coders' notes) was symbolized identically by both coders. A mismatch or disagreement was scored if an action was symbolized differently or missed by one coder. Using this system the frequency pairs (frequency of agreement, frequency of disagreement) for the four process-sequence behaviors were:

DS	(working forward)	(146, 29)
DA	(working backward)	(33, 16)
T	(trial and error)	(41, 4)
↓	(structural error)	(37, 5)

To determine intercoder reliability on the four performance score measures, the frequencies of agreement or disagreement corresponded to the number of observations in which the numerical point scores of both coders in each category were equal or unequal. The respective coefficients for approach score, plan score, result score, and total score were .94, .93, .97, and .96. Thus, the two coders agreed consistently on all aspects of score performance.

Table 4

Classification System

Uses mnemonic notation	Score: result
Representative diagram-yes	Score: total
Representative diagram-no	Reads problem (R)
Recalls related problem	Separates/summarizes data (S)
Uses method of related problem	Draws diagram (M_f)
Uses result of related problem	Diagram with Coordinate System (M_{f_c})
Routine check of manipulations	Synthetic Deduction (DS)
Checks if result is reasonable	Analytic Deduction (DA)
Checks if all information used	Trial and Error (T)
Checks for appropriate dimensions	Reasons by Analogy (An)
Makes algebraic manipulation error	Produces Equation/Relation (Me)
Makes numerical computation error	Algorithmic process (Alg)
Makes differentiation error	Nonclassifiable (N)
Other errors	Checks solution (C)
Time: excluding looking back	Makes structural error (↧)
Time: looking back	Makes executive error (↓)
Time: total	Notices/corrects error (*)
Score: approach	Hesitates; two units (-)
Score: plan	Stops without solution (/)

After making adjustments to the coding scheme, 38 classifications emerged to form the system used in analyzing problem-solving protocols. These are listed in Table 4.

Analysis of Data

Each behavior/event appearing in Table 4, with the exception of the three time scores, was converted to a dichotomous variable for analysis. Depending on symmetry or asymmetry of the distribution, the criterion for dichotomization was set either at the integer nearest the median of a variable's posttest frequency distribution or on the basis of simple presence or absence of the behavior. Frequency scores were obtained by summing observations of the behavior for each subject across the seven posttest problems. Response level 0 meant absence of the behavior or frequency below the median; response level 1 meant presence of the behavior or frequency above the median.

Two types of X^2 analyses were used to treat dichotomous data. First, posttest data were entered in two-way contingency tables for standard X^2 analysis of main effects due to treatment. The data were explored further by entering pretest and posttest information on 14 subjects into three-way contingency tables and applying a method of logit analysis suggested by Goodman (1969, 1970, 1971). This analysis was used to explore potential main effects

and interactions of pretesting condition or pretest response-level with treatment condition and posttest response-level. To illustrate the nature of these interactions, sample three-way contingency tables are given in Tables 5 and 6.

Note that three-way interactions in a three-way table are equivalent to two-way interactions in a 2 × 2 factorial arrangement, with the dependent measure being a dichotomous posttest response level. Analogously, two-way interactions in a three-way table are equivalent to main effects in the factorial perspective.

Nondichotomous data (e.g., time scores) were treated by analysis of variance and covariance using standard *F*-tests (Winer, 1962).

Although the statistical analysis of data described above lends an appearance of hypothesis-testing, this was an exploratory clinical study. Various hypotheses were devised, but the statistical tests were applied primarily to provide insight into which behaviors might be worth pursuing in further work. Six hypotheses formed the framework of the investigation at this point. Five of these hypotheses were concerned with pretest × treatment interaction, main effect due to pretesting, pretest response level, main effect due to treatment; the posttest scores were used as dependent measures for each of these hypotheses. The sixth hypothesis concerned the general effect of treatment on combined posttest scores. Where interactions were shown to exist, main effects were not explored further. However, if the data reflected an apparent difference between experimental and control groups, and if there were no apparent interactive effects of pretesting or pretest response level with treatment, or main effect of treatment upon pretested groups, then the behavior variable under consideraion was regarded as potentially influenced by treatment alone and should merit further experimentation.

The number of subjects in each treatment condition and posttest response level for 35 dichotomous behavior and event variables are given in Table 7.

Table 5

Sample Three-way Contingency Table
(*N*=30)

Pretested				Unpretested			
Experimental		Control		Experimental		Control	
0	1	0	1	0	1	0	1
f_{111}	f_{112}	f_{121}	f_{122}	f_{211}	f_{212}	f_{221}	f_{222}

Note. Pretesting Condition × Treatment × Posttest Response Level (dependent).
f_{ijk} = observed frequency in cell (i,j,k).

Table 6

Sample Three-way Contingency Table
(*N*=14)

Pretested				Unpretested			
Experimental		Control		Experimental		Control	
0	1	0	1	0	1	0	1
f_{111}	f_{112}	f_{121}	f_{122}	f_{211}	f_{212}	f_{221}	f_{222}

Note. Pretest Response Level × Treatment × Posttest Response Level (dependent).
f_{ijk} = observed frequency in cell (i,j,k).

Table 7

Frequency of *S*s in Each Treatment Condition and Posttest Response Level for 35 Behavior Variables

Variable	Heuristic group Posttest response level		Control group Posttest response level			
	0	1	0	1	χ^2	
Mnemonic notation	6	11	12	1	9.997	$p < .005$
Representative diagram (yes)	8	9	5	8	.222	
Representative diagram (no)	10	7	3	10	3.833	$p < .06$
Recalls related problem	10	7	9	4	.041 (Y)	
Uses method of related problem	1	16	6	7	4.617 (Y)	$p < .05$
Uses result of related problem	8	9	12	1	4.904 (Y)	$p < .05$
Routine check of manipulations	9	8	6	7	.136	
Is result reasonable	7	10	9	4	2.330	
All information used	11	6	10	3	.103 (Y)	
Test by dimensions	6	11	7	6	1.033	
Algebraic manipulation error	7	10	6	7	.074	
Numerical computation error	8	9	3	10	.938 (Y)	
Differentiation error	9	8	8	5	.222	
Other errors	8	9	7	6	.136	
Score: approach	2	15	8	5	6.126 (Y)	$p < .025$
Score: plan	5	12	9	4	4.693	$p < .05$
Score: result	7	10	10	3	3.833	$p < .06$
Score: total	5	12	9	4	4.693	$p < .05$
Rereads problem (R)	15	2	5	8	6.126 (Y)	$p < .025$
Separates/summarizes data (S)	4	13	9	4	6.266	$p < .02$
Draws diagram (M_i)	10	7	8	5	.023	
Draws diagram with coordinate system (M_{i_c})	3	14	3	10	.008 (Y)	
Synthetic deduction (DS)	8	9	5	8	.222	
Analytic deduction (DA)	5	12	8	5	3.096	$p < .10$
Successive approximation (T)	12	5	7	6	.314 (Y)	
Reasoning by analogy (An)	13	4	13	0	1.022 (Y)	
Equation/relation (Me)	8	9	8	5	.621	
Algorithmic process (Alg)	8	9	6	7	.002	
Nonclassifiable (N)	13	4	4	9	6.266	$p < .02$
Checking (C)	6	11	7	6	1.033	
Structural error (↕)	11	6	6	7	1.033	
Executive error (↓)	9	8	9	4	.814	
Error noticed and corrected (°)	5	12	7	6	1.833	
2-unit hesitation (-)	12	5	2	11	9.020	$p < .005$
Stops without solution (/)	10	7	8	5	.023	

Y = Yates' correction.

Results

Across all 35 dichotomous variables and five hypotheses dealing with confounding interactive and main effects there were only eight instances in which the X^2 analysis indicated significance at or below a p level of .10. These will be discussed first since the associated data is not part of this paper (see Lucas, 1972).

Two variables, routine checking of manipulations and trial and error (T), exhibited interactive effects of pretesting × treatment ($p < .05$ and $p < .10$, respectively). The pretested control group checked problems very infrequently, while just the opposite was true for the unpretested control group. However, both experimental groups showed little variation on checking across pretesting levels. Pretested experimental subjects used trial and error much less than pretested control subjects, while unpretested groups showed little difference. The most significant information on pretesting × treatment interaction was that generally the novelty of the interview situation and the heuristic mode of instruction did not interact to have a noticeable effect on posttest problem-solving.

The next question explored dealt with main effects of pretesting alone. There seemed to be at least a marginal effect ($p < .10$) in the following three cases: recalling related problems, committing structural errors, and stopping without solution. The data revealed that, if anything, pretesting had a negative effect on recalling related problems. On the other hand, pretesting appeared to have a marginal tendency ($p < .10$) to reduce frequency of structural errors. Subjects who had been pretested were probably more careful when reading problem statements and representing conditions. Pretested subjects showed more persistence in pursuing a problem to completion than unpretested subjects. The latter stopped without solution in 29 instances and the former in only 11.

There were no interactions of pretest level × treatment which seemed to affect posttest problem-solving. Pretested control subjects committed slightly fewer executive errors than the other groups; however, there were indications that the experimental groups noticed and corrected errors more frequently and had higher result scores.

In probing the relationship between pretest and posttest response levels, the following question was posed: "Assuming no pretest × treatment interaction, do those who tend to score high (low) for a given variable on the pretest tend to score similarly for that variable on the posttest?" In response, one variable stood out clearly; that variable was drawing diagrams ($p < .05$). It appeared that the treatment had no influence on this behavior. Further probing of the related checklist categories "representative diagram-yes" and "representative diagram-no" supported the hypothesis that poor

problem solvers draw poor diagrams and good problem solvers draw good diagrams, and that exposure to heuristics has little effect on this disposition.

A time-score analysis of variance was made of the three nondichotomous variables dealing with time measurements. Pretesting alone did not have a significant effect on any time variable. Moreover, pretesting did not interact with treatment to produce effects on posttest time scores. Also, when the pretest time scores were used as covariates to adjust posttest time scores, there was no significant difference on time excluding looking back or on total time, but the data indicate an appreciable difference favoring the experimental group on time spent looking back.

After probing the data for interactions and main effects of pretesting condition, response levels, and treatment upon posttest problem-solving and finding minimal potential influence in most cases, the question of potential main effects due to treatment alone remained. Combining the information from Table 7 with the pretesting data and related tests, the following results were observed.

I. Significant differences attributed to exposure to heuristics were found

1. on *heuristic strategies:*
   using mnemonic notation ($p < .005$)
   using methods of related problems ($p < .05$)
   using results of related problems ($p < .05$)
   separating/summarizing data ($p < .02$)
   reasoning by analogy (marginal, $p < .10$)
2. on *measures of difficulty:*
   rereading the problem ($p < .025$)
   frequency of hesitation ($p < .005$)
3. on *performance scores:*
   approach score ($p < .025$)
   plan score ($p < .05$)
   result score ($p < .06$)
   total score ($p < .05$)

Also, experimental subjects spent significantly more time looking back at a problem ($p < .08$), drew fewer nonrepresentative diagrams ($p < .06$), and exhibited less nonclassifiable behavior (mumbling, unclear statements, manipulations without rationale) ($p < .02$). Slight, but not significant, differences which seemed to favor the experimental group were found on time excluding looking back, checking to see if a result is reasonable, and explicitly correcting errors.

II. No apparent effects of exposure to heuristics were found

1. on *heuristic strategies:*
   drawing diagrams
   representative diagram-yes
   diagrams with coordinate system
   checking (all categories—frequency or nature)
   synthetic deduction (working forward)
   successive approximation
   producing equations (translating conditions)
   reasoning by analogy

2. on *measures of difficulty:*
   stopping without solution

3. on *errors:*
   structural errors
   executive errors (all categories—frequency or nature)

There was also no significant difference observed on total solution time or usage of algorithmic processes, the latter being almost identical for both heuristic groups.

A more detailed discussion of these results can be found in Lucas (1972).

Discussion

The objectives of this study were to explore, conjecture, and generally set a course for continued investigation. At its onset, certain questions were posed as guidelines. These are paraphrased here.

1. Can a suitable instrument be devised for observing and recording process actions and events in problem-solving of college students? If so, is the system reliable? Which heuristics are found in problem-solving among young adults?

2. Can instruction in heuristics effect strategy shifts or influence problem-solving performance? If heuristics can be "taught," what effect are observable and measurable?

3. Can instruction in heuristics be integrated into a standard content-oriented course such as calculus without significantly restructuring the course? If so, do heuristics learned in the context of a particular subject transfer to more general mathematical problems?

The system of behavioral analysis became increasingly important as an end in itself during this study. Kilpatrick's system was revised to delete nonheuristic behaviors and to add a number of heuristics which occurred dur-

ing the pilot study. After several revisions and tests of reliability, a system feasible for observing and recording process behavior emerged. While its application required painstaking effort, it was, at the time, superior to any system known to the investigator for classifying problem-solving strategies. Moreover, the system was moderately reliable given its complexity and the nature and number of judgments to be made.

The investigator regarded the system of behavioral analysis as simply an approximation to one instrument for measuring various dimensions of mathematical problem-solving. Much work needs to be done in the area of instrumentation. This problem has become a major thrust of individual and team research by the investigator.

Of all the behaviors represented in the coding system, inductive reasoning and looking back in search of alternate solutions and invented problems were observed least frequently in the protocols of first-year college students. The investigator guessed that induction, or searching for patterns, may be a problem-specific behavior and whether a subject chooses to look back or not may be a function of the experimental situation. Retrospective comments participants made after interview sessions tended to bear this out. Further work on the looking back heuristic (Smith, 1973) is currently in progress.

In this study, changes in strategy and improved performance were very positive indicators that heuristics can be taught. Students exposed to heuristics approached problems in a more organized fashion. They preferred to use mnemonic notation, their planning was more explicit, they organized problem information more effectively, and while the experimental treatment did not appear to influence diagramming, students exposed to heuristics generally constructed their diagrams more carefully.

Drawing inferences from related problems—that is, building a bridge which connects the given situation with prior information (Wickelgren, 1974)—also appeared to be influenced positively by heuristic instruction. Experimental subjects applied methods and results of related problems more frequently than control subjects. This supported a similar result obtained by Larsen (1960).

The processes of synthesis (working forward) and successive approximation (trial and error) were apparently unaffected by the heuristics instruction, as were translating problem conditions and algorithmic exercises; however, there were indications that emphasis on heuristics did influence reasoning by analysis (working backward). While examining the data and unknown of a problem, experimental subjects were frequently observed breaking the problem down into a sequence of subproblems or subgoals. This heuristic seems to be related to an attitude of explicit planning.

One of the most disappointing results observed throughout the study was the general absence of looking back behaviors. Two related classifications

mentioned earlier (alternate solutions and posing new problems) were dropped early in the study because of low frequency of observation, and in the final analysis there were no significant differences on the nature and frequency of checking. This outcome was puzzling in view of the fact that looking back heuristics had been emphasized and encouraged throughout the instructional period.

The strongest evidence of the influence of heuristics was obtained from certain aspects of general problem-solving performance. Experimental subjects exhibited clearly superior performance on all score attributes—approaching the problem, devising workable plans, obtaining accurate results, and total score. Both groups, control and experimental, were similar on frequency or nature of errors, but the experimental group noticed and corrected errors more often. Also, experimental subjects seemed to have less difficulty with problems. They reread and hesitated much less frequently and they usually started their solutions with greater ease. On the other hand, there was no real difference between groups on stopping without solution, a behavior related to perseverance. Finally, while there was no significant difference on frequency of looking back, experimental subjects spent more time looking back.

The teaching experiment itself demonstrated that heuristics could be integrated into a content-laden curricular structure like university calculus and still have positive effects. In both the pre- and posttests, two of the test problems were calculus problems, and the other five embodied general, non-calculus, mathematical situations. Those subjects trained in heuristics applied their training not only to calculus problems, but transferred it to general problems as well. It was not clear whether parallel instruction in a standard course or central emphasis in a special problem-solving course would be more conducive to learning heuristics. However, the reported study demonstrated the possibility of parallel instruction, and the author has subsequently developed a seminar to explore the effectiveness of special problem-solving courses.

Implications for Research

At the time of this writing there was still an acute need for exploratory teaching experiments aimed at learning and teaching heuristics. Information gained from such studies can provide direction to researchers and provide fresh ideas for the learning and teaching of mathematics.

The study reported here was concerned with many heuristics. Looking back, it would probably be well to limit the scope of similar investigations to only one or a few heuristic behaviors. Having many similar behaviors to separate and make judgments about results in confusion when basic definitions and coding decisions have to be made.

Another limitation of this study was the relatively small number of subjects. However, this design can be defended since the mode of interview and protocol analysis demands small groups, clinical settings, or case studies. It is not feasible to collect data in this manner from large groups without using many additional trained coder-interviewers whose judgments have been cross-checked for reliability. It is also this writer's opinion that problem-solving research, especially the heuristic dimension, has not advanced to the point of employing rigorous experimental designs using large randomly selected groups and associated statistical tests of high power.

Systems of behavioral analysis and instruments for measuring heuristic actions within problem-solving need further refinement and testing. Optimally, a system is needed which is sensitive to heuristics, other problem-solving events, and performance as well as one which is applicable to all problem areas of mathematics and human developmental levels. One implication of this study is that any such system must incorporate the concept of process-sequence. In the study, frequencies of heuristic usage and score measures were used. This was a shortcoming, because other important information is accessible through this system. For example, there are behavior patterns and styles peculiar to individuals and problems. Some subjects initiated a problem solution without any explicit plan, others specified a complete plan at the outset, and still others produced fragments and subplans as the solution progressed. Similarly, some subjects never checked their work, others checked the entire problem after completion, and others checked back after each productive step. These are examples of patterned or stylistic behavior that should motivate further investigation for which elaborate instruments will be necessary.

The lack of inductive reasoning and looking back behaviors raises questions about task variables, problem-specific heuristics, and situation variables. Kilpatrick (1975) has discussed research variables and methodologies quite thoroughly. Questions related to problem-specific heuristics might be: "Do certain problems, by their structure, tend to elicit pattern search behavior?"; "Does the way in which a question is asked influence the heuristic direction a solver will take?"; and "Does interview time, presence of interviewer, or number of test problems inhibit certain heuristics, e.g., looking back?" The study reported here suggests that the thinking-aloud data-gathering method be coupled with retrospection by the solver to maximize the information available about the solution process.

Assuming that heuristics can indeed be taught, the larger question, "How?", still remains with us. Is Polya's system of questions and suggestions sufficient? Must instruction include many mathematical problems, or can instruction using selected problems and sequences yield the same results? Should heuristics be identified explicitly and their application demonstrated in an expository manner, or should the teaching process evolve organically, implicitly, and without labels? Specifically, what actions of the teacher promote

effective use of heuristics in the learner? Experimentation is needed which better controls the teacher variable.

The question of retention of heuristic strategy remains unresolved. Once learned, do heuristics need to be practiced? How long must students be exposed to instruction in heuristics before positive, lasting effects take hold? Do heuristics learned on specific classes of problems transfer to general problems and vice-versa? While this study hints at positive effects after applying heuristics to calculus problems, this investigator has a suspicion that a course in general problem-solving would be more effective.

The reported study involved young adults at the college level. Some of Polya's heuristics were present in their problem-solving and some were not. Further study is needed to uncover relationships between developmental level and learning heuristics. Are there "golden moments" for learning certain heuristics? Are some heuristics never learned unless explicitly taught? What can we do in elementary and secondary schools to promote communication and use of heuristic methods for better problem-solving?

Finally, a process-oriented system of scoring problem solutions needs to be developed (see Kantowski, 1974). What makes a good problem solver? We, as researchers and teachers of mathematics, have an idea of what we mean by a good problem-solver, but we have not specified well-defined characteristics. For example, one problem-solver may exhibit many varied and sophisticated heuristic processes in a correct solution while another may use apparently few heuristics but solve the problem much more quickly. Who is the better problem solver? What is an "elegant" solution? Perhaps we will always have differences of opinion in answering these questions, but it is time we address them squarely.

Mathematical problem-solving and heuristic strategy need much more attention from research. Research efforts in this area are closely linked with the core of teaching and learning in the classroom, for it is through problem-solving that students discover mathematics and it is through heuristics that students discover problem-solving.

Postscript

The research reported here has since been extended by the writer to motivate curriculum development and research in mathematical problem-solving. One very pleasant and challenging outgrowth of this research was the development of a university mathematics seminar specifically concerned with mathematical problem-solving. The course was designed by the author during a 2-year period following the completion of his dissertation, and was offered for the first time in the spring term of 1974 at the University of Wisconsin—Oshkosh. Its basic structure reflects an integration of mathematics and psy-

chology: the heuristic concepts of Polya and others (see Rubinstein, 1974 and Wickelgren, 1974), the instructional format of R.L. Moore (Whyburn, 1970) in which participants present and analyze their problem solutions, and a collection of nonroutine problems drawn from various branches of mathematics at an elementary level. The chief objective of this seminar is to improve communication of mathematical problem-solving, with enhanced individual problem-solving as a potential byproduct. It is clear to the writer that a one-semester course in problem-solving probably has little effect on long-term strategy shifts in the problem-solving of adults. However, getting preservice and inservice teachers of secondary school mathematics conversant with questions, suggestions, alternate solutions, patterns, analogies, and heuristic strategy is a step in the right direction.

Unlike most mathematics courses, this seminar offers problems selected with the intent of reinforcing heuristic strategy rather than specific mathematical concepts. The problems are the medium of instruction, while the "concepts" are heuristic techniques. Consequently, participants are encouraged to talk about their problem solutions, especially their strategies. Each solution is critiqued by the group; the solver must defend not only his or her mathematical statement, but also the reasoning behind it.

The seminar in mathematical problem-solving has provided fertile ground for cultivating ideas on teaching and learning mathematics; it also has motivated extended research on heuristics. The writer has developed new perspectives on heuristics by observing and discussing the problem-solving of advanced undergraduate and graduate students. The use of symmetry as a tool in solving problems, the transfer of a technique from one branch of mathematics to another (for example, Polya's level lines strategy in analytic geometry and Lagrange multipliers in analysis), and the distinction between Pappus's working backward technique and Wickelgren's (1974) subgoals method are examples of these new perceptions.

In spring 1975, the Georgia Center for the Study of Learning and Teaching Mathematics (GCSLTM) called together 45 researchers in mathematical problem-solving for a conference on research issues and ideas. The primary thrust of this and several other GCSLTM conferences was an attempt to consolidate research in mathematics education. The writer was invited to the problem-solving conference, which had the theme "Heuristics." At the conference, current variables, models, and methodologies for research were summarized (see Hatfield, 1975 and Kilpatrick, 1975) and a report on "teaching experiment" research in the Soviet Union was given (Kantowski, 1975b). A general report of GCSLTM research activities was presented to the International Congress on Mathematical Education at Karlsruhe, West Germany in the summer of 1976 (Hatfield, 1976).

One of the tangible outcomes of the GCSLTM problem-solving conference was the formation of research teams around shared areas of interest for

collaboration on common researchable problems. Currently, the writer is a member of a six-person team conducting clinical studies on heuristics at various developmental levels. Adult participants from the writer's problem-solving seminar have served as subjects for a small-scale teaching experiment modeled after the dissertation study reported here, but including more heuristics and a modified system of analysis. Over a 2-year period this team has concentrated on developing a team-constructed, reliable, effective system of behavioral analysis. Part of this effort includes developing an adequate system for scoring problem solutions. A report of the team's activities was made to the National Council of Teachers of Mathematics at their national convention in Cincinnati in spring 1977. A monograph is in press at the time of this writing.

During the next decade, collaborative efforts by researchers using the mathematics classroom as a laboratory should yield effective models for learning and teaching heuristics at all levels.

Chapter 6

A Multidimensional Exploratory Investigation of Small Group-Heuristic and Expository Learning in Calculus

Norman J. Loomer

The research reported in this chapter explored the effects of two different methods of calculus instruction—the small group-discovery method pioneered by Davidson (see Chapter 3) and an expository method. In this study carefully prescribed models of teacher behavior for the two methods and an inventory, called the Measure of Teacher Fidelity to the Model, were developed. The inventory records students' perception of the teacher's classroom behavior. The models and inventory will make extension of this research easier, since they enable an investigator to show that the instruction was administered as prescribed.

The author taught two intact college Calculus I classes, one by each method, for one semester. He selected criteria for a multidimensional comparison of the two methods, and selected and developed instruments to measure instructional outcomes along those criteria. Instructional outcomes were measured at three points during the study: Observation 1 occurred immediately before the instructional phase, Observation 2 took place at the end of the instruction, and Observation 3 occurred 1 month after Observation 2, immediately following a college vacation. This chapter reports the results of the evaluation and analyzes the evaluation procedures and instruments. The study was exploratory, not experimental. It was only a first step toward bridging the large gap between Davidson's (1971a) original feasibility study and a large-scale, carefully controlled, experimental study. The statistical analyses, therefore, do not yield firm conclusions but rather define and sharpen hypotheses about the differential effects of the two methods. The results must therefore be regarded as tentative.

Discovery teaching is sometimes assumed to be synonymous with heuristic teaching, but Higgins (1971, p. 487) observes that to a mathematician the word, *heuristic* has an infinitely richer meaning than simply *discovery*. He urges that a teaching technique be called heuristic if it (a) approaches content through problems, (b) reflects problem-solving techniques in the logical construction of instructional procedures, (c) demands flexibility for uncertainty and alternate procedures, and (d) seeks to maximize student action and participation in the teaching-learning process. (p. 494). Believing that the small group-discovery method meets these four criteria, the author has taken the liberty of renaming it the *small group-heuristic method.*

Method

Instructional Procedures

Many investigations of teaching methods have been criticized because quite different methods are labeled the same. Moreover, the methods compared are seldom described in behavioral terms that are precise enough to allow replication (Fey, 1969, pp. 536, 545; Richards, 1973, p. 149; Tanner, 1969, p. 654; Wittrock, 1966, pp. 44-45). To avert such criticism of the present study, models of rigorously defined patterns of teacher behavior for the small group-heuristic method and an expository method of teaching calculus were developed.

The models grew out of discovery-expository research in elementary school arithmetic by Worthen (1968, pp. 225-227) and Robertson (1970, pp. 30-38). Modifications of the discovery teaching model were made to make it appropriate for the small group-heuristic method of teaching calculus and were strongly influenced by the instructional practices and classroom organization employed by Davidson (1971a, pp. 100-102, 162-166) and the heuristic questioning style advocated by Polya (1957).

The models for the two methods are differentiated along eight dimensions: (a) organization of the class, (b) initiation of learning experiences, (c) interjection of teacher knowledge, (d) questioning and answering procedures, (e) appraisal techniques, (f) control of student interaction, (g) use of instructional materials, and (h) determination of policies. Brief summaries of model teacher behavior for each method along these dimensions follow.

Small Group-Heuristic Method

Organization of the class. The learning unit is a group of three or four students. Within each group the students learn together by doing problems, exploring questions, and proving theorems.

Initiation of learning experiences. Exploration precedes formalization. The teacher initiates student exploration of a concept mainly by raising questions. If the teacher chooses to formalize the concept, he or she waits to do so until the group has successfully completed its exploration.

Interjection of teacher knowledge. The teacher tries to develop a learning climate that permits students to show their knowledge, and therefore does not act as the primary source of information. Instead suggestions are given for solving problems only when asked or when help is clearly needed. Even then the teacher encourages students to contribute to the solution. He or she is receptive to different approaches to stimulate students' ideas and suggestions.

Questioning and answering procedures. When asking or answering questions the teacher avoids giving too much information. He or she tries to ask questions or indicate steps that could have occurred to the students themselves. Questions should apply not only to the problem at hand, but also,

whenever possible, to other similar problems. Many times these questions come from Polya's "How to Solve It" list (1957, pp. xvi-xvii).

Appraisal techniques. If an atmosphere for exploration is to prevail, the teacher must be sensitive about appraising student responses. He or she does not judge incorrect responses in a negative manner, but uses them to stimulate a continued search for a solution. If students are unsure of a response, the teacher may encourage a guess or hunch. Students are urged to find their own errors by using Polya's "Looking Back" heuristics. (Polya, 1957, p.14).

Control of student interaction. The teacher encourages group members to work together cooperatively, and to build the ideas of others to achieve group solutions. He or she discourages interaction between groups to avoid interfering with each group's opportunity to achieve its own solution.

Use of instructional materials. A group is led to discovery of a mathematical concept by exploring problems prepared by the teacher and distributed to each student.

Determination of policies. By class discussion and decision-making by majority vote, the students determine policies on grading, scheduling of examinations, the manner of forming work groups, and standards of behavior for the work groups.

Expository Method

Organization of the class. The learning unit is the entire class. The students learn mathematics by watching the teacher, asking questions, responding to questions, and doing daily homework problems.

Initiation of learning experiences. Formalization precedes exploration. The teacher states definitions, proves theorems, and describes concepts before exploring them by means of examples. Then the students can explore them in homework problems.

Interjection of teacher knowledge. The teacher acts as the primary source of mathematical knowledge, by indicating to students that he or she will always be able to work a problem correctly if they cannot.

Questioning and answering procedures. The teacher asks the class questions that are simple, close-ended, and directed specifically to the concept being discussed. He or she immediately recognizes incorrect answers, gives students an opportunity to correct their own mistakes, and responds to student questions by reiterating a principle or relationship. The teacher may use an example to clarify the way a principle or relationship is used to solve a problem.

Appraisal techniques. The teacher shows great concern for errors. He or she follows a student's incorrect responses with a discussion on why the

errors are incorrect, takes care not to judge students negatively and warns students about common errors and uses examples to emphasize them.

Control of student interaction. The teacher encourages students to share their ideas about a problem with the class.

Use of instructional materials. The teacher uses the textbook, which has expository characteristics, as the primary source of materials and ideas.

Determination of policies. The teacher determines virtually all policies, including the method of grading and scheduling of exams and quizzes.

In addition to the eight dimensions along which the models for the two methods differ, there are three dimensions along which they coincide: sufficient teacher preparation, teacher enthusiasm, and nonevaluative climate. Brief summaries of model teaching behavior along each dimension follow.

Sufficient teacher preparation. The teacher has teaching materials ready at the beginning of each class.

Teacher enthusiasm. The teacher projects a sincere enthusiasm for mathematics, for the students, and for the teaching method that he or she is using.

Nonevaluative climate. The teacher does not make value judgments when responding to students and establishes a climate in which they are free to respond even when uncertain of their answers.

Criteria for the Comparison

Many research designs used to compare two treatments have been criticized for attempting to show one method superior to another by measuring on a single criterion, usually achievement (Begle & Wilson, 1970, p. 368; Fey, 1969, p. 536; Shulman, 1970, pp. 36-37). This approach is insensitive to the differential effects of the two methods. In the present study, therefore, instructional outcomes were measured by multiple criteria.

A review of the literature on expository-discovery research (Brown, 1971; Fey, 1969; Hughes, 1974; Scott & Frayer, 1970; Shulman, 1970;Tanner, 1969; Willoughby, 1969; Wittrock, 1966) revealed that the enthusiasts for discovery learning claim that it enhances motivation and retention of concepts, and develops problem-solving ability, the heuristics of discovery, deeper understanding of concepts and structure, and realistic insights into how mathematics grows. Those less enthusiastic about discovery learning counter that it is too time-consuming, offers little to the learner that cannot be offered by good expository teaching, and is not beneficial for the cognitively sophisticated individual.

Consideration of these assertions led to the selection of the following criteria for comparative evaluation of the small group-heuristic and expository methods of teaching calculus: (a) calculus achievement, (b) calculus achieve-

ment at the computation-comprehension cognitive level, (c) calculus achievement at the application-analysis cognitive level, (d) mathematical problem-solving achievement, (e) mathematics attitudes, (f) problem-solving behaviors, (g) retention of calculus achievement, (h) retention of calculus achievement at the computation-comprehension cognitive level, (i) retention of calculus achievement at the application-analysis cognitive level, (j) retention of mathematical problem-solving achievement, (k) retention of problem-solving behaviors, and (l) rate of coverage of material in each method.

Tests and Measures

Measures of Teacher Fidelity to the Model. To assess the degree to which the teacher adhered to the models, the investigator adapted an inventory developed by Worthen (1968). Called the Measure of Teacher Fidelity to the Model, it was administered to the students in the calculus classes after the instructional phase of the study. The inventory consists of a series of statements drawn from the models, about teacher behavior to which the student responds "A" if the teacher *almost always* did it in the class, "B" if the teacher *often* did it, "C" if *sometimes*, "D" if *seldom*, and "E" if the teacher *almost never* did it.

The statements are of three types: those that refer to the expository-heuristic characteristics of the method, those that refer to characteristics on which the methods are to coincide, and, among the statements drawn from the model for the small group-heuristic method, those that refer to operation of the small groups. The three types of items make up an Expository-Heuristic Scale, a Coinciding Characteristics Scale, and a Small Group Operation Scale. The 34 items on the Expository-Heuristic Scale are classified into items that, if answered affirmatively, typify student perception of highly expository teaching behavior (E items) and those that typify highly heuristic behavior (H items). The five items on the Coinciding Characteristics Scale are classified into items that typify desirable teaching behavior (DT items) and those that typify undesirable teaching behavior (UT items). The six items on the Small Group Operation Scale, to which only students in the small group-heuristic class responded, are classified into items that typify desirable group operation (DG items) and those that typify undesirable group operation (UG items).

Six sample items, one from each classification, are listed below.

5. Our teacher _____ created the feeling that a stigma was attached if a student made an error. (UT)

18. Our teacher was _____ enthusiastic toward mathematics, toward the students, and toward the method of teaching he was using. (DT)

19. Our teacher _____ gave us a rule or procedure to use for solving new kinds of problems. (E)

33. When our teacher assisted us in the solution of a problem, he _____ gave suggestions that applied not only to the problem at hand but also to problems in general. (H)

41. In our class the members of each group _____ worked together cooperatively to achieve a group solution to a problem. (DG)

44. Our teacher _____ encouraged members of a group to talk to members of another group or observe the activities of another group. (UG)

A response to an item is scored as follows: On the H, DT, and DG items, the response is assigned a value of 4 for almost always, 3 for often, 2 for sometimes, 1 for seldom, and 0 for almost never. On the E, UT, and UG items the scale is reversed. On each inventory an index between 0 and 100 is obtained for each scale by multiplying the mean of that scale's item scores by 25.

The following four tests were administered as shown on Table 1 to gather data about student achievement, attitude, and behavior in each of the experimental classes:

Calculus Achievement Tests. The Calculus Achievement Tests include multiple-choice items drawn from the 1969 Advanced Placement Examination in Mathematics (Finkbeiner, Neff, & Williams, 1971), sample Advanced Placement Examinations (College Entrance Examination Board, 1972), and items developed by the investigator. Three judges, including the investigator, classified each item by content category and National Longitudinal Study of Mathematics Abilities (NLSMA) (Romberg & Wilson, 1969) cognitive level. For the purpose of this study the four NLSMA cognitive levels were compressed into two, Computation-Comprehension and Application-Analysis. Each item was then entered into a content category by cognitive level matrix, and from each cell items were randomly assigned to two forms of the Calculus Achievement Test. Because not all topics originally scheduled were actually covered by both calculus classes, several items were eliminated from the scoring, resulting in abbreviated forms called Form A* and Form B*.

Problem-solving achievement tests. Problem-solving achievement was defined to be the score on a multiple-choice problem-solving tests. The problems were to be what Dodson (1972, pp. 3-6) calls "insightful"—not routine textbook problems but problems requiring original thinking. A wealth of these problems was found on the Preliminary Contest Examinations of the Wisconsin Section of the Mathematical Association of America, offered annually in Wisconsin high schools "to discover and encourage talented students" (Buck, 1959, p. 202). Problems from the examinations were classified by difficulty level and content category—algebra, geometry, or number systems—

Table 1

Administration Schedule for Criterion Measures

Observation 1	Observation 2	Obervation 3
Calculus Achievement Test, Form A*	Calculus Achievement Test, Form A*	Calculus Achievement Test, Form B*
Mathematical Problem Solving Test, Form A	Mathematical Problem Solving Test, Form B	Mathematical Problem Solving Test, Form C
Mathematics Attitude Scale	Mathematics Attitude Scale	
Problem Solving Attitude Scale	Problem Solving Attitude Scale	
Problem Solving Interview	Problem Solving Interview	Problem Solving Interview
	Measure of Teacher Fidelity to the Model	Measure of Teacher Fidelity to the Model

and items from each cell of the resulting matrix were randomly assigned to three forms of the Mathematical Problem Solving Test. Responses to items on both this test and the Calculus Achievement Test were scored 4 if right, -1 if wrong, and 0 if omitted.

Mathematics attitude measures. Romberg (1969, p. 481) argues that a single, global measure of attitudes toward mathematics is not realistic, since there is probably a set of feelings that vary from computation to problem-solving. The assessment of attitudes in this study, therefore, had three phases: a measure of general attitudes toward mathematics; a measure of attitudes toward problem-solving, which was an important variable in the two instructional methods; and an open-ended questionnaire to elicit the reaction of the small group-heuristic class to their teaching method. The measure of general attitudes toward mathematics was Aiken and Dreger's (1961) "Revised Math Attitude Scale," which was called the Mathematics Attitude Scale during this study. It consists of 10 statements connoting negative attitudes and ten statements connoting positive attitudes toward mathematics, to which the student responds to one of five Likert alternatives. The Problem Solving Attitude Scale was constructed by the investigator, who selected 16 statements specific to problem-solving, eight positive and eight negative, from attitude-toward-mathematics instruments developed by Coon (1969, pp. 175-176), Cummins (1958, pp. 179-181), and Worthen (1965, pp. A3.30-A3.31). For each statement the student responds to one of five Likert alternatives. The Small Group Calculus Class Questionnaire is an adaptation of the one developed by Davidson for his feasibility study. The student responses for each question were classified into various categories and counted.

Procedures for assessing problem-solving behaviors. Student problem-solving behaviors were assessed using the diagnostic procedure developed by Kilpatrick (1967) and Lucas (1972) (also see Chapter 5). Students participated in 1-hour interviews during which they thought aloud as they solved three mathematical word problems. The interviews were tape recorded, and the taped commentaries and written work were used as the basis for analysis using Kilpatrick's and Lucas's system of behavioral analysis. The system used in this study evaluates 59 aspects of problem solving activity, which fall into five categories: (a) heuristic strategies, (b) modes of difficulty, (c) types of errors, (d) performance measured by time, and (e) performance measured by score.

Procedures for the Pilot Study

The pilot study of the two instructional methods was conducted during the fall semester of 1973-1974 at Ripon College, a small, private, coeducational, liberal arts college in east central Wisconsin. Two classes of Calculus I were offered, both taught by the investigator, for which students registered in the usual way. Students were therefore not assigned randomly to the two classes, but neither were they selected in any special way. It was expected that

there would be no significant initial differences between the two classes on any of the selected criteria and that the statistical design would adjust for any minor differences. Teaching methods were assigned randomly to classes.

Because teacher interaction with the small groups would be an important and time-consuming activity in the small group-heuristic class, enrollment was limited to 16 students. Enrollment in the expository class was not limited: 25 students registered, and 16 of them were randomly selected to participate in the evaluation phases of the study.

There were three observations of student performance: Observation 1 at the beginning of the semester immediately before the instructional phase, Observation 2 at the end of the semester upon completion of the instructional phase, and Observation 3 one month after Observation 2, immediately following a college vacation.

Table 1 lists the measures administered at each observation. For Observation 1 all measures except the problem-solving interviews were administered during the first three class meetings. The interviews were conducted in the investigator's office during the first 8 days of the semester. The Observation 2 interviews were conducted during the last 8 days of the semester, and with the exception of the Calculus Achievement Test the remaining measures were administered during the last two class meetings. The Calculus Achievement Test was administered during the final examination at the same time to students in both classes. The Observation 3 interviews were conducted during the first 8 days of the second semester. The remaining measures were administered during a special testing session on the day before that semester's classes began.

Subjects

Most of the students participating in the evaluation phases of the study were 18-year-old male freshman mathematics or science majors with 8 semesters of high school mathematics and mathematics grade point average above 3.00. Most of them had not studied calculus in high school, had ACT Mathematics scores of at least 28 or SAT Mathematics scores of at least 600, and had a high school percentile rank of at least 90.

Attrition during the study reduced the number of subjects from 16 to 13 in each class: Two students in the small group-heuristic class were unable to attend testing sessions because of illness and one declined to participate in the last two problem-solving interviews. Three subjects in the expository class withdrew near the end of the semester because of unsatisfactory grades.

Instructional Materials

In the expository class, the textbook *Calculus of One Variable, Second Edition* by Seeley (1972) was used. In the small group-heuristic class the instructional materials were based largely on prepublication materials for the book *Calculus: A Student Discovery Approach* by Davidson and Leach (1973). The author found it necessary, however, to revise these materials be-

cause the content did not match the topics to be covered in Calculus I, and in some respects the organization did not conform to the author's biases about how the ideas of calculus should be developed. In his revisions the author was guided by the following principles: (a) the need for processes should be established before teaching them; (b) the concrete should precede the abstract; (c) the approach to concepts should be intuitive rather than rigorous; (d) definitions and symbols should be introduced only after the student has had extended experience with the ideas they represent; and (e) "Let us teach proving by all means, but let us also teach guessing" (Polya, 1963, p. 606).

Statistical Design

Three statistical models were used to analyze the data gathered during the study: analysis of variance for data on three measures of academic status prior to instruction (high school percentile rank, SAT Mathematics score, Ripon College Mathematics Placement Test score), for Observation 1 data on the Calculus Achievement Test, Problem Solving Achievement Test, Mathematics Attitude Scale, and Problem Solving Attitude Scale, and for data from the Measure of Teacher Fidelity to the Model; analysis of covariance for Observation 2 and 3 data on the Calculus Achievement Test, Problem Solving Achievement Test, Mathematics Attitude Scale, and Problem Solving Attitude Scale and for three measures of time taken during problem-solving interviews; and logit analysis (Goodman, 1970) for the analysis of 56 other measures of problem-solving behaviors. Potential covariates for each performance measure subjected to analysis of covariance included the three measures of academic status prior to instruction, the Observation 1 scores on that performance measure and on the Calculus Achievement Test, Mathematical Problem Solving Test, Mathematics Attitude Scale, and Problem Solving Attitude Scale. The covariates for each performance measure were selected using a step-by-step regression procedure described by Draper and Smith (1966, pp. 171-195).

Results

Significance Levels

For meaningful interpretation of data on instructional outcomes, it was crucial to find clear evidence that the two classes received instruction that differed consistently on the expository-heuristic characteristics of the instructional model and agreed consistently on the coinciding characteristics. It was therefore important not to infer a difference on the expository-heuristic characteristics of instruction when one did not exist (i.e., make a Type I error) and important not to fail to infer a difference on the coinciding characteristics of instruction when one did exist (i.e., make a Type II error). Accordingly, for analysis of the data on the expository-heuristic characteristics of instruction the significance level was set at .01 to minimize the chance of a Type I error, and for analysis of the data on the coinciding characteristics of instruction it was set at .10 to minimize the chance of a Type II error.

The purpose of analyzing the data on instructional outcomes was not to make generalizations, but to probe for conjectures to serve as the basis for future experiments. Thus the objectives of the study were threatened more by failure to infer a treatment effect when one did exist (Type II error) than by inference of a treatment effect when one did not exist (Type I error). Accordingly, for analysis of the data on instructional outcomes the significance level was set at .10 in order to minimize the chance of Type II error.

Analysis of Initial Data

In order to determine the comparability of the two calculus classes prior to instruction, data on the initial measures were subjected to analysis of variance. Table 2 shows that, contrary to the investigator's expectations, the calculus classes were not equivalent at the beginning. On every measure the mean of the expository class exceeded the mean of the small group-heuristic class, and on two measures, the SAT Mathematics Test and the Mathematical Problem Solving Test, the difference in means was statistically significant ($p < .10$).

The nonequivalence of the calculus classes on these measures subjects the results of the analysis of covariance to interpretation difficulties. First, the covariance adjustment may not have removed all bias; some bias may be present from a disturbing variable that was overlooked. Second, when the covariates showed real differences between the groups, covariance adjustments involved extrapolation. Consequently, the farther apart the groups were on the covariate means, the more imprecise was the estimate of the difference in the adjusted means. Thus the adjusted differences may be insignificant statistically because the adjusted comparisons are of low precision (Cochran, 1957, pp. 265-266). Therefore, interpretation of the results of the analysis of covariance may be speculative.

Analysis of Teacher Behavior

Data from the Measure of Teacher Fidelity to the Model at Observation 2 provide clear evidence that the teacher taught the two classes in close conformity to their respective models. Table 3 shows the means by class for each of the three scales of the fidelity measure. The means on the Expository-Heuristic Scale show that the students perceived the teacher's behavior to be heuristic in the small group-heuristic class and expository in the expository class; analysis of variance indicates that the difference in means is highly significant ($p << .001$). Analysis of data from the Coinciding Characteristics Scale by analysis of variance yields a nonsignificant difference in means ($p > .10$) indicating that, as desired, the students did not perceive the teacher's behavior to differ on the coinciding characteristics of the model. The mean of the scores of the small group-heuristic class on the Small Group Operation Scale has a 90% confidence interval of (77.8, 88.9), indicating that the

Table 2

Analysis of Variance for Initial Measures

Measure[a]	Means		*df*	*F*
	Heuristic	Expository		
High School Rank (100)	83.2	94.2	1/24	2.20
SAT Mathematics Test (800)	621.8	689.8	1/24	5.63*
Ripon Mathematics Placement Test (108)	50.8	60.1	1/24	1.15
Mathematics Attitude Scale (80)	57.9	65.1	1/24	2.26
Problem Solving Attitude Scale (64)	42.3	45.6	1/24	0.77
Calculus Achievement Test, Form A* (96)	2.8	6.7	1/24	0.77
Computation-Comprehension Scale (52)	1.8	3.8	1/24	0.53
Application-Analysis Scale (44)	1.0	2.8	1/24	0.73
Mathematical Problem Solving Test, Form A (52)	10.5	22.9	1/24	4.36*

[a]Numbers in parentheses indicate maximum possible score.

*$p < .05$.

Table 3

Analysis of Teacher Behavior Data

Scale	Small group-heuristic class		Expository class		*df*	*F*
	Mean	Perfect score for model	Mean	Perfect score for model		
Expository-heuristic	75.2	100	24.5	0	1/24	302.56*
Coinciding characteristics	90.0	100	95.0	100	1/24	1.56
Small group operation	83.2	100	—	—	12	—

*$p << .001$.

students perceived small group operation to be in close conformity to the model.

Selection of Covariates

Table 4 lists the covariates selected for each of the instructional outcome variables by the step-by-step regression procedure. The procedure selects only variables that are significantly related ($p < .10$) to the outcome variable. An unexpected discovery was that scores on Form A of the Mathematical Problem Solving Test were significantly related not only to the scores on Forms B and C, as would be expected, but also to five of the six calculus achievement measures. In this study the Mathematical Problem Solving Test was a better predictor of calculus achievement than the measures usually used for this purpose—high school rank, SAT mathematics score, and placement examination score. The use of a problem-solving test in predicting calculus achievement appears to be an important subject for further investigation.

Analysis of Instructional Outcome Data

The evidence from the teacher behavior data makes possible meaningful interpretation of the data obtained to compare instructional outcomes on the 12 selected criteria. Table 5 shows the results of analysis of the data for criteria 1-5, which concern outcomes measured immediately after the instructional period.

Calculus achievement. On Form A* of the Calculus Achievement Test at Observation 2 the means, adjusted by the analysis of covariance for initial differences between the two classes, favored the expository class, but the difference did not approach the significance level of .10 chosen for the criterion measures.

Calculus achievement at the computation-comprehension cognitive level. On the Computation-Comprehension Scale of Form A* of the Calculus Achievement Test at Observation 2, the adjusted means slightly favored the expository class, but the difference did not approach significance.

Calculus achievement at the application-analysis cognitive level. On the Application-Analysis Scale of Form A* of the Calculus Achievement Test at Observation 2, the adjusted means favored the expository class, but the difference did not approach significance.

Mathematical problem-solving achievement. On Form B of the Mathematical Problem Solving Test at Observation 2, the adjusted means favored the expository class, but the difference again did not approach significance.

Mathematics attitudes. On neither the Mathematics Attitude Scale nor the Problem Solving Attitude Scale at Observation 2 did the difference in adjusted means approach significance. On the Small Group Calculus Class Questionnaire, administered to students in the small group-heuristic class, the reactions to the method ranged from hostile to enthusiastic. On the negative

Table 4

Covariates Selected for Instructional Outcome Measures

Outcome measure	Covariates selected	
	Observation 2	Observation 3
Attitude measures		
Mathematics attitude	Mathematics attitude	
Problem-solving attitude	Problem-solving attitude	
	Calculus achievement	
Achievement Measures		
Calculus achievement	Mathematics problem-solving	Mathematics problem-solving
		Mathematics attitude
Computation-comprehension	SAT mathematics	Mathematics problem-solving
Application-analysis	Mathematics problem-solving	Mathematics attitude
	Application-analysis	Mathematics problem-solving
Mathematics problem-solving	Mathematics problem-solving	Mathematics problem-solving
	Calculus achievement	
	SAT mathematics	
Heuristic time score measures		
Time: Excluding looking back	Time: Excluding looking back	None
Time: Looking back	Ripon mathematics placement	Time: Looking back
		Ripon mathematics placement
Time: Total	Time: Total	None

Table 5

Analysis of Covariance for Observation 2 Achievement and Attitude Measures

Instructional Outcome Measure[a]	Observed means		Adjusted means		*df*	*F*
	Heuristic	Expository	Heuristic	Expository		
Calculus Achievement Test, Form A* (96)	22.7	34.5	26.2	31.0	1/23	0.66
Computation-Comprehension Scale (52)	17.6	23.1	20.1	20.6	1/23	0.02
Application-Analysis Scale (44)	5.1	11.5	6.9	9.7	1/22	0.98
Mathematics Problem-Solving Test, Form B (52)	12.5	16.9	14.3	15.1	1/21	0.07
Mathematics Attitude Scale (80)	55.5	62.2	58.4	59.3	1/23	0.06
Problem-Solving Attitude Scale (64)	41.7	43.9	42.8	42.8	1/22	0.00

[a]Numbers in parentheses indicate maximum possible score.

side, most students were concerned about covering enough material and were bothered by not having a textbook. A few students thought that the class was less stimulating than others and decreased their interest in mathematics. On the positive side, most students enjoyed doing problems every day, thought that the teacher was effective in giving hints, and thought that the class was more stimulating than others. Several said that the class increased their interest in mathematics.

Problem-solving behaviors. Table 6 contains the X^2 statistics and significance levels for the 56 heuristic variables that were dichotomized for logit analysis of their frequency (or score) distributions. The logit model analyzes the data for the posttest, taking into account the student's performance on the pretest at Observation 1. Its function with respect to qualitative data is analogous to the function of analysis of covariance with respect to quantitative data. No results are reported for one variable because a preliminary analysis indicated an interaction between instructional method and pretest response level, making interpretation of results about main effects due to instructional method doubtful.

The only variable for which the X^2 statistic indicated a main effect is Rereads Problem. The contingency table for the variable revealed that the small group-heuristic method produced a greater tendency to reread parts of a problem than the expository method.

Table 7 shows the results of analysis of covariance of three measures of time (in 15-second units) taken during the problem-solving interviews. The statistics indicated one significant difference: The small group-heuristic class spent more time looking back at the problem and solution after obtaining a result.

Table 8 shows the results of analysis of the data relating to criteria 7-10, which concern retention.

Retention of calculus achievement. At Observation 3, one month after the instructional period, the adjusted means on Form B* of the Calculus Achievement Test favored the small group-heuristic class, a reversal of the result at Observation 2. The difference, however, did not approach significance.

Retention of calculus achievement at the computation-comprehension cognitive level. The adjusted means on the Computation-Comprehension Scale of Form B* of the Calculus Achievement Test showed another reversal from Observation 2, with data at Observation 3 favoring the small group-heuristic class. The difference, again, did not approach significance.

Retention of calculus achievement at the application-analysis cognitive level. The adjusted means on the Application-Analysis Scale of Form B* of the Calculus Achievement Test showed yet another reversal from Observation 2,

Table 6

Logit Analysis for Dichotomized Heuristic Variables at Observation 2

Variable	χ^2	Variable	χ^2
Restates problem	2.18	Comparison with known result	0.00
Mnemonic notation	1.24	Condenses/outlines process	0.00
Representative diagram—yes	1.11	Tries to derive differently	0.63
Representative diagram—no	1.11	Variation by analogy	0.00
Auxiliary lines	1.20	Variation by changing conditions	0.00
Isolates focal points	0.94	Algebraic manipulation error	1.92
Recalls related problem	0.49	Numerical computation error	0.06
Uses method of related problem	0.00	Differentiation error	0.00
Uses result of related problem	0.75	Other executive error	0.69
Inductive reasoning	0.18	Misinterprets data	4.36
Routine check of manipulations	0.14	Misinterprets question	0.59
Is result reasonable?	0.45	Other structural error	3.12
All information used?	0.00	Score: approach	1.76
Test for symmetry	0.00	Score: plan	1.45
Test of dimensions	0.18	Random trial and error	0.00
Specialization	0.59	Systematic trial and error	0.62
Score: result	2.53	Reasoning by analogy	0.00
Score: total	2.73	Not classifiable	2.53
Reads problem	0.00	Checks the result	0.19
Rereads problem	4.99*	Varies the process	0.63
Separates/summarizes data	1.87	Varies the problem	0.00
Draws diagram	—[a]	30-second hesitation	0.85
Modifies diagram	0.36	Stops without solution	4.40
Draws diagram with coordinate system	0.46	Structural error	2.02
Model by means of equation	0.33	Executive error	0.62
Algorithmic process	2.24	Corrects error	3.27
Exploratory work with data	0.00		
Deduction by synthesis	3.12		
Deduction by analysis	4.29		

[a]Results not reported because preliminary analysis indicated an interaction between instructional method and pretest response level.

*$p < .10$.

Table 7

Analysis of Covariance for Observation 2 Heuristic Time Score Variables

Variable	Observed means		Adjusted means		*df*	*F*
	Heuristic	Expository	Heuristic	Expository		
Time: Excluding looking back	150.7	137.8	153.9	134.8	1/22	0.78
Time: Looking back	7.8	1.2	8.7	0.4	1/22	7.42*
Time: Total	158.4	139.1	160.5	137.1	1/22	1.11

*$p < .025$.

Table 8

Analysis of Covariance for Observation 3 Achievement Measures

Instructional outcome measure[a]	Observed means		Adjusted means		*df*	*F*
	Heuristic	Expository	Heuristic	Expository		
Calculus Achievement Test, Form B* (84)	11.2	17.5	16.3	12.4	1/22	0.48
Computation-Comprehension Scale (44)	8.5	11.1	10.7	8.9	1/23	0.30
Application-Analysis Scale (40)	2.7	6.4	4.8	4.2	1/22	0.04
Mathematical Problem-Solving Test, Form C (52)	12.5	27.2	15.3	24.3	1/23	6.62*

[a]Numbers in parentheses indicate maximum possible score.

*$p < .025$.

favoring the small group-heuristic class at Observation 3. The difference, however, did not approach significance.

Retention of mathematical problem-solving achievement. The analysis of scores on Form C of the Mathematical Problem Solving Test showed a dramatic, but curious, shift in problem-solving achievement during the vacation period between Observations 2 and 3. While at Observation 2 the difference between adjusted means did not approach significance, at Observation 3 the difference was highly significant ($p < .025$), favoring the expository class.

Retention of problem-solving behaviors. Table 9 contains the X^2 statistics and significance levels for the 56 dichotomized heuristic variables at Observation 3. The only significant difference gives independent confirmation to the shift in problem-solving achievement detected by the Mathematical Problem Solving Test. On both the total score awarded for solution of the problems and on the subscore awarded for correctness of the results, the expository class was favored. The difference indicated at Observation 2 on the variable Rereads Problem did not persist until Observation 3.

Table 10 shows no significant differences on the three measures of time taken during the problem-solving interviews at Observation 3. The *F*-statistic for Time: Looking Back, which was significant at Observation 2, falls just short of the critical value of 2.96 for significance ($p < .10$) at Observation 3.

Rate of coverage of material in each method. Like most expository-discovery studies, this one found discovery learning to be slower. Of 34 topics scheduled to be covered in Calculus I, the small group-heuristic class failed to cover six; the expository class covered not only all the topics scheduled but also four optional topics.

Analysis of the Evaluation Instruments

Reliability coefficients for the instruments used in this study are in Table 11. The investigator chose 0.90 as an acceptable level for reliability coefficients on the achievement measures and 0.80 as an acceptable level on the attitude and teacher behavior measures. The reliability coefficients for the Expository-Heuristic Scale of the Measure of Teacher Fidelity to the Model, the Mathematics Attitude Scale, and the Problem Solving Attitude Scale were all acceptable. Form A of the Mathematical Problem Solving Test has a reliability coefficient of 0.82, which approaches acceptability. The remaining instruments — Coinciding Characteristics Scale, Small Group Operation Scale, Forms B and C of the Mathematical Problem Solving Test, and Forms A* and B* of the Calculus Achievement Test — had unacceptable reliability coefficients. Before these instruments are used in a large-scale experimental study, their reliabilities must be improved. Suggestions for adding or improving items, along with item analyses and other information regarding the validity of

Table 9

Logit Analysis for Dichotomized Heuristic Variables at Observation 3

Variable	χ^2	Variable	χ^2
Restates problem	2.61	Comparison with known result	0.00
Mnemonic notation	0.40	Condenses/outlines process	0.00
Representative diagram—yes	2.03	Tries to derive differently	3.13
Representative diagram—no	2.03	Variation by analogy	0.00
Auxiliary lines	2.30	Variation by changing conditions	0.00
Isolates focal points	0.25	Algebraic manipulation error	0.40
Recalls related problem	0.00	Numerical computation error	0.35
Uses method of related problem	0.00	Differentiation error	0.49
Uses result of related problem	0.18	Other executive error	2.25
Inductive reasoning	2.20	Misinterprets data	4.42
Routine check of manipulations	3.58	Misinterprets question	0.41
Is result reasonable?	—[a]	Other structural error	0.48
All information used?	0.00	Score: approach	2.06
Test for symmetry	0.00	Score: plan	4.56
Test of dimensions	0.75	Random trial and error	0.00
Specialization	0.19	Systematic trial and error	0.67
Score: result	5.23*	Reasoning by analogy	0.00
Score: total	5.05*	Not classifiable	0.30
Reads problem	0.00	Checks the result	1.44
Rereads problem	0.85	Varies the process	3.13
Separates/summarizes data	1.53	Varies the problem	0.00
Draws diagram	2.98	30-second hesitation	1.03
Modifies diagram	1.12	Stops without solution	1.52
Draws diagram with coordinate system	1.56	Structural error	2.97
Model by means of equation	0.48	Executive error	0.06
Algorithmic process	1.97	Corrects error	1.41
Exploratory work with data	0.00		
Deduction by synthesis	1.73		
Deduction by analysis	1.51		

[a]Results not reported because preliminary analysis indicated an interaction between instructional method and pretest response level.

*$p < .10$.

Table 10

Analysis of Covariance for Observation 3 Heuristic Time Score Variables

Variable	Observed means		Adjusted means		*df*	*F*
	Heuristic	Expository	Heuristic	Expository		
Time: Excluding looking back	164.4	176.6	164.4	176.6	1/23	0.17
Time: Looking back	8.9	2.0	7.6	3.3	1/21	2.82
Time: Total	173.3	178.6	173.3	178.6	1/23	0.03

Table 11

Reliability Coefficients for Instruments Used in the Pilot Study

Instrument	Reliability coefficients	
	Hoyt	Test-Retest
Measure of Teacher Fidelity to Model		
Expository-Heuristic Scale	0.97	0.98
Coinciding Characteristics Scale	0.55	0.64
Small Group Operation Scale	0.52	0.77
Mathematics Attitude Scale	0.95	0.94
Problem-Solving Attitude Scale	0.89	
Mathematical Problem-Solving Test, Form A	0.82	
Mathematical Problem-Solving Test, Form B	0.60	
Mathematical Problem-Solving Test, Form C	0.64	
Calculus Achievement Test, Form A*	0.74	
Computation-Comprehension Scale	0.65	
Application-Analysis Scale	0.54	
Calculus Achievement Test, Form B*	0.75	
Computation-Comprehension Scale	0.58	
Application-Analysis Scale	0.52	

the instruments, may be found in the original report of this study (Loomer, 1976, pp. 200-211, 247-250).

Discussion

This study is an exploratory probe of the differential effects of the small group-heuristic and expository methods of teaching calculus. The nature of the study was to explore the experimental method and to sharpen hypotheses, not to reach general conclusions beyond the particular classes and teacher that participated.

The appropriate warnings having been issued, it is possible to make some observations and conjectures. The clearest evidence of the study is that the teaching methods used in the two calculus classes were faithful to their respective models. However, there were few measurable differences in instructional outcomes. Immediately after instruction there were no statistically significant differences between the two classes on any of the attitude or achievement measures. Analysis of 59 heuristic variables produced only two significant differences: The small group-heuristic class reread parts of a problem more frequently and spent more time looking back at the problem and solution.

One month after instruction the expository class showed a surprising superiority in problem-solving achievement, although it had not been immersed in a problem-solving environment during the instructional phase as had the small group-heuristic class. The statistical analyses detected no other significant differences between the two classes on calculus achievement measures or other heuristic variables. There was faint evidence of a reversal on all three calculus achievement measures over the college vacation. The adjusted means, which all favored the expository class at Observation 2, all favored the small group-heuristic class at Observation 3. None of the differences even approached significance, however.

The lack of differences in instructional outcomes despite clear evidence of differences in teaching methods is puzzling. Small sample sizes, low reliabilities of some of the evaluation instruments, or the nonequivalence of the classes prior to instruction may have decreased the precision of the statistical tests. A more carefully controlled, large-scale study using improved instruments would have a better chance of detecting differences.

The results suggest that the small group-heuristic method was much less effective in producing changes in problem-solving behaviors than Lucas's inquiry method, which emphasized instruction in heuristics (see Chapter 4). The key to the difference in the results of the two studies may be the quantity of instruction in heuristics. Lucas was able, as Polya suggests, to make continual use of heuristic questions and suggestions in the classroom. The present investigator was able to make use of heuristic questions and suggestions less frequently — only when a group asked for his help in solving a problem.

Furthermore, he was unable to make a systematic presentation of heuristic suggestions; the heuristic strategy discussed at a particular moment was the one needed by the group at that time.

Two of the instruments developed for this study appear worthy of further study and evaluation. The success of the Measure of Teacher Fidelity to the Model in detecting differences in the expository-heuristic characteristics of the teacher's behavior indicates the possibility of designing other such instruments. The Mathematical Problem Solving Test may be a good instrument for predicting achievement in calculus.

Further exploration of the small group-heuristic method in a large-scale experimental study now seems in order. This exploratory study has laid the groundwork: It has selected criteria and developed instruments for evaluating the method and measuring instructional outcomes, developed models of teaching behavior and an instrument for measuring fidelity to the models, and generated and sharpened hypotheses about the effects of the method.

Chapter 7

A Study of Problem-solving Performance Measures

Donald L. Zalewski

Purpose

One of the primary goals of school mathematics programs is to develop problem-solving abilities. Helping students develop these abilities and assessing their problem-solving performance are joint concerns of curriculum, instruction, and research. At present the most appropriate way to assess students' problem-solving performance seems to be through the use of personal interviews and the thinking aloud procedure (Kilpatrick, 1967; Loomer, 1976; Lucas, 1972). However, these techniques are very time consuming and cannot easily be employed by school mathematics teachers. The study reported in this chapter involves the development and testing of a paper-and-pencil instrument intended to predict a student's level of problem-solving performance.

Definitions

The content of any problem-solving study depends on its interpretation of the term "problem." In this study, a *mathematical problem* is one which meets three conditions: (a) the statement presents information and an objective or question whose answer is based on that information; (b) the objective or answer to the question can be found by translating the information into mathematical terms or by applying results from mathematics; (c) the individual attempting to answer the question or attain the objective does not possess an immediate answer, procedure, or algorithm which solves the problem. If an individual solved a given problem or one similar to it previously and simply recalls the answer or procedure, the situation would not be considered a problem for that person. *Mathematical problem-solving* is the process of developing and using a procedure to solve a mathematical problem. The process involved may require a search among possible strategies, the use of various rules and techniques, and prior knowledge of mathematics.

Background

Commercial Instruments

While developing test items for a state mathematics assessment program, the investigator realized that very few methods exist to record and assess

the mathematical problem-solving achievement of students. During the initial part of this study, the investigator found a few procedures which claim to measure problem-solving achievement, but an examination of these procedures raised doubts about their validity.

Commercial tests include the mathematical problem-solving measures which are most available for school use. However, as the problem-solving subtests were examined, several inadequacies in the items and scoring procedures were detected.

The Iowa Test of Basic Skills (Lindquist & Hieronymus, 1964), Form 2, is identified as a problem-solving assessment instrument. However, the items do not satisfy the definition of a mathematical problem used in this study because direct algorithmic processes are suggested by words such as "total" and "difference." "Mathematical problem solving" is one of the tests in the Metropolitan Achievement Test (Durost, Bixler, Wrightstone, Prescott & Balow, 1970) batteries, but the items are simple verbal situations. They require only one obvious operation suggested by questions such as "How much more. . . ?," "How many times as many . . .?," or "What is the area of . . . ?" In items calling for two operations and more complex solving behaviors, the students need only select the appropriate sentence from four choices (the fourth being "more information needed") without actually solving the problem.

The Instructional Objectives Exchange (IOX, 1970) identifies a major category, "Application—Problem Solving." The questions give attention to both process and solution, but the sample objectives emphasize the answers to the items and a student is rated only on the number correct.

The California Achievement test battery (Tiegs & Clark, 1970) uses a "Problems" test which allows 13 minutes to solve 15 written items about money, averages, area, volume, and percents. Eight of the problems require only one operation and seven items require two operations. Scoring is based only on the number of correct responses.

All the commercial tests the investigator examined give a choice of answers (usually four or five) for each item and score a student according to the number of correct choices. Though this practice permits rapid scoring, it does not create a genuine problem-solving situation.

The validity of the commercial tests as problem-solving measures became even more questionable as the investigator examined their validation procedures. A search of both the technical and teacher's manuals of the Iowa Test of Basic Skills (ITBS) battery (Lindquist & Hieronymus, 1964) failed to uncover any validation procedures for their "problem-solving" test (A-2). The writers' statement, "The most valid achievement test for your school is that which in itself defines most adequately your objectives of instruction,"

seems to summarize their attitude toward test validity, especially in the area of mathematical problem solving.

The Metropolitan Achievement Test manual (Durost, Bixler, Wrightstone, Prescott, & Balow, 1971a, 1971b) discusses test validity, but fails to provide a definition of mathematical problems or any interpretation of test results in terms of problem-solving skills. The writers state that the content validity of this test was established by examining textbooks, study guides, and mathematics curriculum recommendations. The teacher's handbook offers advice similar to that of ITBS, "Since each school has its own curriculum, the content validity of Metropolitan Achievement Tests must be evaluated by each school." (p. 32) Construct validity is concerned with "the completeness of the test as a well rounded or representative sample of the content we are hoping to measure, and also the appropriateness of the types used." (p. 32) The test writers believe that concurrent validity and predictive validity have little or no meaning as applied to specific tests within achievement batteries and no validity measurements are offered.

The content validity of the California Achievement test battery is discussed very briefly; it was based on widely accepted mathematics curriculum objectives in the United States.

The examination of commercial tests as mathematical problem-solving measures revealed several reasons to doubt their validity: (a) the "problems" were usually simple written items (often referred to as "word" or "story" problems) which did not meet this study's definition of mathematical problems; (b) the scoring only focused on the correct response without considering the processes used; (c) the tests set time limits which gave students little opportunity to practice problem-solving techniques; and (d) the test writers provided no validity measures except the usual content validity statements. Thus, the commercial tests were judged not to be valid mathematical problem-solving measures and other procedures were examined.

Research Procedures

Research in problem solving has been hampered by semantic ambiguities, overgeneralizations, and lack of consolidation of efforts; however, some helpful directions and procedures have resulted. It is generally agreed that the products of problem solving — responses, results, or completed methods—do not permit sound inferences about the processes used and that it is necessary to study subjects' observable behaviors to better analyze problem-solving practices. Several procedures of varying utility and validity have been devised to generate and record an observable sequence of behavior. Bourne and Battig (1966) described a sample of frequently employed methods and commented on their limitations. For example, manipulative devices such as pendulum problems (Maier, 1931) or jars of water (Luchins & Luchins, 1950) only revealed a few of the hypotheses or hunches a subject was entertaining at a

given moment. The limitations of attempting to infer process from external actions made the direct exploration of mental processes a desirable alternative.

The direct investigation of problem-solving processes requires subjects to verbalize during or after the solution search. *Introspection* requires a subject to solve problems and report on thoughts, reactions, and feelings while performing. Though introspection externalizes thought patterns, there are serious questions about the distortion and interference introduced by the experimental procedure. *Retrospection* requires the subject to give a narrative account of his or her thoughts and processes after having completed the problem-solving task. Broder and Bloom (1950) found that when this procedure was used to osbserve problem-solving tasks some steps were forgotten and rearrangement of the remaining steps in a more logical order resulted. In addition to these internal deficiencies, the two verbalization techniques are expensive in time and equipment, and require careful training of both subjects and observers.

One method that avoids some of these difficulties is the thinking aloud technique in which the subject simply verbalizes (without analyzing) thoughts while working, and these statements are recorded. The thinking aloud method has been criticized and questioned, but evidence concerning whether speech and thinking complement or interfere with each other has been inconclusive. Kilpatrick (1967) was willing to risk these possible dangers in return for the helpful information that can be gained.

> The method of thinking aloud has the special virtues of being both productive and easy to use. If the subject understands what is wanted—that he is not only to solve the problem but also to tell how he goes about finding a solution — and if the method is used with the awareness of its limitations, then one can obtain detailed information about thought processes. (p. 8)

The increasing recognition and use of the thinking aloud procedure in research studies provided sufficient reason to assume that the procedure was valid for identifying problem-solving behaviors. The patterns and processes revealed by subjects' responses during taped pilot study interviews added to the investigator's confidence that the thinking aloud procedure reflects genuine problem-solving behaviors.

A complication of the thinking aloud procedure is that the recorded verbal data has to be analyzed and classified. During his investigation of eighth graders' problem-solving performance, Kilpatrick (1967) devised a general guide for coding audiotaped protocols of students thinking aloud while solving mathematical problems. Subsequently, he devised a comprehensive system which included a checklist and a model for coding the chain of behaviors occurring in a subject's protocol.

Lucas (1972) extended Kilpatrick's classification system in a study involving heuristic problem-solving strategies in calculus. He altered the checklist, added symbols, made numerous revisions in the process coding system, and developed a scoring system based on performance within a problem. Although Lucas used his revised coding scheme to detect changes in heuristic solving strategies in calculus, the form is easily adaptable to any study involving mathematical problem solving in thinking aloud interviews.

Goals of This Study

The investigation of research procedures reported here found a method that was assumed to be valid and reliable for recording and assessing mathematical problem-solving behaviors. The first goal of this study was to record, assess, and rank the mathematical problem-solving performances of seventh-grade students using the thinking aloud procedure and Lucas's refined coding system. Two questions related to this goal were considered:

1. How well does the thinking aloud procedure and coding scheme capture and classify the mathematical problem-solving behaviors of seventh-grade students?

2. Is it possible to separate and rank seventh-grade students according to their coded problem-solving protocols?

Analysis and evaluation of the coded data from thinking aloud sessions was assumed to be a valid method of classifying students' mathematical problem-solving performances. However, the method is not readily used in schools because of its physical limitations: Only one subject can be tested at a time; considerable time and expense are involved in recording, coding, and evaluating each performance; and specially trained interviewers and coders are needed. These factors would make a large scale school assessment financially impractical, if not impossible. For an individual teacher, the lack of interview and coding skills could be a handicap, and finding additional time for interviews in an already crowded schedule makes the time required for the thinking aloud procedure a deterrent for classroom use.

A practical alternative for measuring problem-solving performance is a paper-and-pencil instrument, as it requires only simple materials and need not be administered by specially trained personnel. The second goal of this study was to investigate the feasibility of producing a written instrument that reflects the mathematical problem-solving ability of seventh graders. Specifically, the question being asked was, "Is it feasible to construct a written evaluative instrument whose results correlate well with the ranking derived from the coded protocols?"

Study Design

This study had three principal parts. First, the problem-solving performances of students observed thinking aloud were recorded, analyzed, and ranked. Second, a written test (WT) was devised and administered to the same students to provide a second ranking. Third, the correlation between the two ranks was determined. Details of each part of the study will be discussed separately.

Part I: The Complex Problem-solving Assessment Procedure

In planning to use the complex interview and coding procedure, the mathematical problems, subjects, interviews, coding system, and ranking procedures all received careful scrutiny. These considerations will be described in turn.

The mathematical problems. Six interview test (IT) items were drawn from a pool developed by the investigator using the definition of "mathematical problem" given earlier. The judgments of mathematics educators, the results of a pilot study, and an examination of the mathematics curriculum provided in textbooks were used to screen items and strengthen content validity. To prevent computational difficulty from being an important factor, the arithmetic in the problems was kept simple.

The subjects. The seventh-grade level was chosen for this study. A single grade was chosen to restrict the scope of the study and to extend the work started by the investigator during the state assessment of seventh-grade students.

The interviews. Seventh-grade students solved six mathematical problems in a thinking aloud taped interview. The interview procedures, developed by Kilpatrick (1967) and Lucas (1972), are detailed in the author's dissertation (Zalewski, 1974).

The coding system. The coding system for this study was a combination of Kilpatrick's (1967) and Lucas's (1972). Lucas developed a five point scoring system based on a subject's complete protocol for a problem. He totaled the points for Approach (0 or 1), Plan (0, 1, or 2), and Result (0, 1, or 2). Lucas's scoring procedure was followed as the IT ranking of students was developed. Lucas's system is a modification of Kilpatrick's, but since Lucas used his system to code the behavior of calculus students, some symbols and items were eliminated. Other revisions were made according to the results of a pilot study.

The ranking. Two measures were applied after coding and scoring the subjects' protocols. The number of correct answers was the simplest measure while the total process score (or any of the subscores) provided a second basis for ranking subjects. Both statistics were considered in determining the students' IT rankings.

A third basis for ranking interview subjects was a statistical analysis of their coded protocols. Latent partitioning (Lord & Novick, 1968; Torgerson, 1958) or a type of clustering analysis (Hubert, 1973) was applied. Based on patterns of the coded behaviors, these statistical procedures provided a separation of the subjects into subgroups. Then the investigator determined an ordering between and within subgroups to provide yet another ranking of the IT subjects.

Part II: The Written Test

In this part of the study a paper-and-pencil instrument (WT) was devised to provide a second ranking of the subjects who participated in the Part I interviews. Subjects took the WT and were ranked according to the results. The correlation between the rankings of Parts I and II were established statistically in Part 3. A high correlation would provide the concurrent validity needed to suggest that substituting a written test for the complex interview and coding procedure is feasible.

The WT items. It was desired that the WT be related to mathematical problem solving. Thus, the items chosen for the paper-and-pencil instrument were mathematical in nature, nonroutine, and open-ended. Manipulations, symbols, number size, and number of steps were kept within the ability of seventh graders.

The WT items were not the same as the mathematical problems used in the IT. Some items in the WT require only one-step solutions and did not meet the criteria of mathematical problems, but all attempted to avoid simple recall of knowledge. An item pool was created in accord with these criteria; it was randomly sampled in generating the WT. For convenient school use, the WT was constructed so that it can be administered to students in one 50-minute class period.

The WT ranking. On the WT, the subjects were ranked solely on the basis of correct responses. The answer to a problem alone does not reveal the solution processes involved, but the WT was not designed to measure processes. Its only purpose was to provide a second ranking of the same students who were given the IT.

Part III: The Comparison of Ranks

The third part of this study was designed to test similarities between the rankings developed in Parts I and II.

Correlations. After the rankings from Parts I and II were established, Spearman's rank order correlation coefficient and Kendall's τ (1955) were computed. A correlation of at least .71 would indicate the WT scores account for approximately 50% of the variance in the IT ranks and establish concurrent validity of the WT. This was determined to be the minimum correlation

to support the feasibility of using the WT as a substitute for the thinking aloud and coding procedure.

The Studies

Prior to the main study, a pilot study was conducted; that study resulted in important changes in the main study. Thus, both studies are reported here.

Pilot Study

The purpose of the pilot study was to tryout the interview procedures and their coding and scoring schemes and to use an initial version of the WT. The pilot study results suggested changes in the original plan for the study and modifications were made in the taping format, the WT length, the interview procedures, and the checklist and coding scheme.

Audiotaping versus videotaping. During the summer of 1973, eight volunteers who had completed seventh grade in Madison, Wisconsin, took both the WT and the IT. After audiotaping the verbalizations of the first subject, it was apparent that interesting physical actions and silent indications of problem-solving processes were not being captured. For example, a subject moved his pencil across the page as he silently reread; the audiotape recorded only silence while this significant behavior occurred. The investigator decided to use videotaping with four pilot subjects to explore the advantages of a visual and audio record of the interviews. Later the use of videotape was incorporated into the main study.

During the pilot study, it seemed that pilot subjects who were videotaped behaved differently than if they had been audiotaped. Thus a question arose: Do subjects perform differently if they are videotaped instead of being audiotaped? To answer this, two measures of difference based on problem-solving interview scores were compared through a one-way fixed effects analysis of variance (ANOVA) with the subjects randomly assigned to treatment groups (audiotaping or videotaping). The following hypothesis was posed:

> Hypothesis H1: The mean score on achievement for videotaped subjects equals the mean score on achievement for audiotaped subjects.

An arbitrary significance level of .05 was chosen for rejection of this null hypothesis.

A second one-way fixed effects ANOVA was applied to the total time each subject used to solve the six mathematical problems given during the interviews. A second hypothesis with a .05 rejection level was posed:

> Hypothesis H2: The mean solution time of the videotaped subjects equals the mean solution time of the audiotaped subjects.

The incorporation of videotaping into the study evoked one issue which was not directly related to the data. Lucas (personal communication) coded the protocols obtained during the pilot tryout in this study and observed that it took noticeably less time to code videotaped protocols than audiotaped protocols. To explore this difference systematically, each taping procedure was considered a treatment, and subjects were randomly assigned to permit an ANOVA. The hypothesis tested was the following:

> Hypothesis H3: The mean coding time for audiotaped protocols equals the mean coding time for videotaped protocols.

An arbitrary significance level of .10 was chosen for rejection of this hypothesis.

Because a statistically significant difference in coding times may not be important in practice, a second method of comparing coding times was planned. The difference between the average coding time for 1 minute of audiotape and the average coding time for 1 minute of videotape would be found; if the difference between the averages was greater than 10%, that difference would be regarded as significant.

Changes in the WT. The original written test contained 16 items. For this test, Hoyt's internal consistency measure produced a reliability of only 0.1765. This extremely low reliability could have been due to the small number of subjects in the pilot study, an unusual interaction of subjects and items, or the number of items on the test. It was assume that the first two possibilities would be compensated for in the main study by the larger number of subjects and the random item sampling procedure. The third possible cause of low reliability was counteracted bby increasing the WT from 16 to 20 items.

Changes in the interview procedures. The pilot study produced two changes in the interview strategies. First, the apparent nervousness and haste of pilot subjects who were videtaped suggested that extra efforts would have to be made to put students at ease before having them think aloud while solving problems. Subsequently, the interviewer planned to verbally emphasize that the subjects could use as much time as they needed, would converse with each subject until the student appeared comfortable, and would not place a clock in a conspicuous position. The same precautions were planned for audiotaped subjects although the presence of a tape recorder did not seem to have the same effect as a camera.

The second change in the interview procedures resulted from following Lucas's (1972) practice of verbally encouraging a subject to think aloud if the student fell silent for a period of 30 seconds. When one subject was prodded with "What are you doing now?" after a silence of 30 seconds, he appeared slightly irritated at having his thoughts interrupted, replied "I'm thinking," and lapsed back into silence. Similar reactions by other subjects persuaded the investigator to avoid interfering after 30 seconds and to use his discretion if the

subject was not talkative or overly active for more than a minute, especially if it appeared that the subject was stymied or frustrated. The interviewer would not interrupt a subject if it appeared that he or she was silently devising a plan, even though this neglect would cause gaps in the thinking aloud record of a student's problem-solving procedures.

Changes in the coding system and checklist. In addition to the changes in the interview procedures, some modifications of Lucas's (1972) coding system were suggested by the pilot study. The subjects in the pilot study never produced behaviors to be coded as M_{f_c} (introducing diagram with coordinate system imposed), Vs (varies the process), or Vm (varies the problem). These symbols and the related items on the checklist were eliminated from Lucas's (1972) format. Additional symbols were devised to classify behaviors which did not fit easily into Lucas's system: Rs (restates the problem in his or her own words), Rr (rereads the problem or parts of it), DX (exploratory work with data), TR (irregular trial and error), and Ts (systematic trial and error). The changes in the process symbols were accompanied by modifications in the items on the checklist.

Main Study

The main study was conducted according to the modified plans resulting from the pilot study. The IT, WT, population, and events are described below.

IT and WT. After creating a pool of 50 representative mathematical problems, the investigator randomly selected six items for the IT. The WT was created by randomly selecting 20 items from the pool of 165 items described earlier.

Population. The study was conducted at an elementary, parochial school located in west central Madison, Wisconsin. Its 435 first- through eighth-grade students came mainly from middle to upper middle class families of white color workers and professionals. The mathematics program in grades 5-8 was partially individualized, and students worked at their own pace.

Written test administration. The two seventh-grade mathematics teachers administered the WT to all 63 seventh graders. Each class had approximately 40 minutes to complete the test with extra time allowed for those students who needed it. The procedures for administering the WT did not directly follow the plans. Originally, half of the subjects would have taken the WT after the IT, but school conditions dictated otherwise and all subjects took the WT before the IT.

Another change in plans occurred in the WT item format. Originally, the 20 items were to be presented in random order to each student to avoid a sequence effect. This arrangement would have required that each of the 63 tests be typed individually. To permit rapid production of the WT, the origi-

nal plan was abandoned and all 20 items were presented in the same order to every subject.

After the written tests were completed, the investigator visited the classrooms to discuss the WT with the subjects and to seek their cooperation in arranging the thinking aloud interviews. All subjects were encouraged to participate whether or not they believed they had done well on the WT. Subjects were not told their results on the WT.

The interview sample. To heed Kilpatrick's (1967) concern for the pressure placed upon subjects in interview situations, subjects with at least average mathematical ability were chosen for the interviews. No recent achievement test scores were available to classify students, so before the WT the two mathematics teachers were asked to identify students in their classes who were at least average in achievement. Thirty-one average or above average subjects were identified; all of these students accepted an invitation to participate in the interviews.

The interview arrangements. The videotaped interviews were scheduled for the last week in February 1974, and the audiotaped interviews were scheduled for the next week. Sixteen of the 31 subjects were randomly selected to be videotaped. The videotaped interviews were conducted in a mobile unit parked beside the school. The 15 audiotaped interviews were conducted in a meeting room in the school basement.

Data and Analyses

This section reports the data from each of the three principal parts of the study. First the scores, rankings, and statistics for the written test will be described. The data from the interview test, the statistical analysis of the relationship between rankings, and the results of exploratory statistical procedures follow.

The Written Test (WT)

The purpose of the WT was to produce an initial ranking of the subjects; they were also to be ranked by their performance on the IT, the mathematical problem-solving instrument. The data and statistics for the WT and a subsequent WT2 are presented before feasibility factors are reported.

Subject response data. A total of 63 seventh-grade students took the 20 item WT. The descriptive statistics for the WT are presented in Table 1 for the 31 subjects who had been rated as average or above average in mathematics achievement (Group A) and the 32 students rated below average (Group B) by their mathematics teachers.

Table 1

Mean, Standard Deviation, and Range for the WT: Group A, Group B, and Combined

	Number of subjects	Mean	Standard Deviation	Range (20 items)
Group A	31	7.4194	3.8796	2 to 14
Group B	32	3.7500	2.7238	1 to 12
Groups A and B combined	63	5.5556	3.7963	1 to 14

According to Table 1, the results on the WT were consistent with teacher ratings. Group A averaged 7.42 correct responses, almost twice the 3.75 mean of the lower rated Group B. Group A omitted an average of 2.7 items on the WT while Group B subjects omitted an average of 4.1 items.

The low mean scores and the high number of items omitted by both groups of WT subjects caused the investigator to question whether the mathematical abilities of the 63 seventh-grade students who participated were representative. In order to compare the subjects to other seventh-grade students, a second 20-item written test (WT2) was developed from the available pool with the restriction that any item which appeared on the WT could not be used on the WT2. In May 1974, 350 seventh-grade students from Madison and Des Moines, including the original 63 from Madison, were given the WT2. The mean for the 63 Madison subjects on this second test was 6.11; this was close enough to the overall mean of 5.93 to assure the investigator that these were typical seventh-grade students and that their low mean scores were due to the general difficulty of the items.

WT length and reliability. The low mean scores of the students did not affect the feasibility of the WT, but two other factors, test length and reliability, were also important. A test which took more than an hour to complete or which did not attain a reliability of .80 would not meet the expectations of the investigator.

Hoyt reliabilities (Hoyt, 1941) were calculated for both the WT and WT2. When the scores of both Group A and Group B were used, the Hoyt reliability of the WT is .82; the reliability of that test is .7968 for Group A alone and .73 for Group B alone.

Using the scores of all 350 students who took the WT2, the Hoyt reliability of this instrument is .84. The corresponding reliabilities for the WT2 when Group A ($N = 31$) only was used and Group B ($N = 32$) only was used were .77 and .68, respectively. No Hoyt reliability was calculated for the WT2 using only the scores from Groups A and B together.

The calculated reliabilities demonstrate that, overall, both the WT and WT2 exceed the reliability level sought. Using only the scores of Group A, the

potential IT subjects, the reliabilities of these two tests are close to the desired level of .80, but when Group B alone is considered, the reliabilities of both WT and WT2 fall short. However, since Group B did not participate in the interview phase of this study, the overall reliabilities are satisfactory, and the overall reliabilities for Group A are near the desired level, it was feasible to compare the results of this test to the problem-solving scores derived from the thinking aloud interviews.

To see if the test was an appropriate length, the investigator recorded completion times for 59 of the 63 WT2 subjects. Mean completion time for these subjects was 27 minutes, and the range was from 16 to 37 minutes. The 27 minute mean indicated that seventh-grade students could respond to the 20 items in one class period. Even subjects taking 15 minutes more than the mean test time would finish the WT in 42 minutes, a completion time less than the maximum 50 minute period.

Written test rankings. The rank of a subject on the WT was based solely on the number of correct responses, and only subjects from Group A who participated in the IT were ranked. Since two written tests, the WT and WT2, were administered, rankings were determined for each and are presented in Table 2.

As can be seen in Table 2, the rankings developed from the WT and WT2 are similar. They agree perfectly on subjects 8 (rank 6.5), 16, and 31, and agree closely on subjects 2, 10, and 27. Despite the high apparent ranking agreement, the investigator decided to compare each WT ranking to the IT ranking separately to see which test produced a stronger relationship.

The Interview Test

Group A, the students designated as being average or above average achievers in mathematics, participated in an interview test (IT) using the thinking aloud procedure. Their problem-solving protocols were coded, scored, and ranked; these data are reported next.

The thinking aloud procedure. During the thinking aloud interviews, the investigator observed four behaviors which might raise questions about the effectiveness of this procedure. The behaviors were subjects' remarks concerning their ability to think aloud, periods of silence, use of retrospection, and subject anxiousness. Table 3 summarizes the occurrences of these behaviors in the videotaped and audiotaped interviews.

As seen in Table 3, two subjects from each taping group made direct comments about their ability to think aloud. For example, subject five worked calmly but quietly, and after reading an IT problem, explained to the investigator, "I'm gonna figure this out in my mind and tell you when I'm done—or else I can't get it."

Table 2

Rankings of Group A Based on the Results of the WT and the WT2

Subject number[a]	WT number correct	WT rank[b]	WT2 number correct	WT2 rank[b]
1	8	14	12	6.5
2	9	11	10	10.5
3	7	16	9	13
4	9	11	12	6.5
5	9	11	7	18.5
6	7	16	4	27
7	5	21	4	27
8	12	6.5	12	6.5
9	6	18.5	5	24
10	13	4	14	2.5
11	5	21	3	29
12	3	27	6	21.5
13	3	27	2	30.5
14	3	27	7	18.5
15	11	8	13	4
16	4	24	5	24
17	13	4	7	18.5
18	9	11	8	15.5
19	14	1.5	12	6.5
20	4	24	11	9
21	9	11	9	13
22	7	16	9	13
23	2	30	5	24
24	2	30	2	30.5
25	12	6.5	16	1
26	13	4	10	10.5
27	14	1.5	14	2.5
28	4	24	6	21.5
29	5	21	4	27
30	2	30	8	15.5
31	6	18.5	7	18.5

[a] The subject number represents the order of his or her appearance in the interviews. Subjects 1 to 16 were videotaped and subjects 17 to 31 were audiotaped.

[b] In case of ties on number correct, the ranks were averaged.

Table 3
Indicators of Thinking Aloud Difficulties

	During videotaping	During audiotaping
Number of subjects who made comments on their thinking aloud ability	2	2
Number of subjects who explained by retrospection	5	4
Number of silent pauses which occurred:		
30 to 60 seconds	20	25
over 60 seconds	19	21
Number of subjects who were judged to be anxious	7	6

Retrospection was used by subjects who explained their procedures after they had achieved an answer. Five videotaped subjects practiced retrospection in a total of 10 instances with one subject resorting to retrospection on all five of the problems she solved. Four audiotaped subjects accounted for eight instances of retrospection.

Silent pauses were periods of time when subjects produced no codable behavior while attempting to solve a problem. Pauses of less than 30 seconds were often used for assimilating information, organizing ideas, or silent recapitulation and were not considered to indicate thinking aloud difficulty. However, pauses longer than 30 seconds usually occurred in protocols of subjects who had difficulties expressing their thoughts aloud. All pauses over 30 seconds were recorded and dichotomized: pauses less than 1 minute and those longer than 1 minute. As indicated by Table 3, silent pauses occurred frequently in both types of taping.

The last category in Table 3 records subjects' unspoken reactions while participating in the interviews. Four videotaped subjects and three audiotaped subjects were clearly nervous. The most common and obvious signs included tapping a pencil, scratching parts of the body, or frequent shifting of body positions. Three other subjects from each taping procedure exhibited less obvious nervous behaviors such as reading the problems rapidly or carelessly and sometimes slurring or mispronouncing words.

The coding system. During the pilot study, the investigator was fortunate to receive Lucas's personal assistance in checking the application of his system. Calculating a direct ratio of the frequency of agreement to the total frequency of agreement and disagreement between Lucas and the investigator, acceptable agreement measures were computed for the process-sequence coding (.72), the checklist (.67), and the scoring system on Approach (.93), Plan (.86), and Result (.86). However, the modifications of Lucas's (1972) system for this study necessitated additional agreement measures. Three coders, among them the investigator, were used to establish those agreements. The resulting agreement-disagreement ratios produced an agreement measure of

.83 across all variables and interjudge reliability tests produced a measure of .80.

After agreement ratios and reliability measures were computed and evaluated, the coded protocols and scores were used to search for ranking schemes.

The IT ranking schemes. Application of Lucas's (1972) scoring system produced four measures for each problem: Approach (0 or 1), Plan (0, 1, or 2), Result (0, 1, or 2), and Problem Total (0-5). The first ranking scheme (Ranking A) was developed by summing the six Problem Totals for each subject and assigning a rank of 1 to the highest sum and a rank of 31 to the lowest sum. Tied ranks were averaged. The totals and ranks for Ranking A are presented in Table 4 as are those for Rankings B and C.

According to Ranking A, subject 15 had the highest total interview test score (24 points) and was ranked first, while subjects 24 and 29 scored no points and shared the last averaged rank of 30.5. Other ties occurred at totals of 18, 10, 9, 8, 5, 4, and 3 points. Five subjects tied at 9 to share a rank of 14 (average of 12-16) and five other subjects tied at 8 to share rank 19 (average of 17-21). Except for three subjects tied at 18 points, the remaining ties occurred in pairs.

The large number of ties in Ranking A made it likely that this ranking would produce a low association with written test ranks. Thus Rankings B and C were developed to differentiate between subjects. Subjects with tied totals earned different numbers of points in subscores of the scoring system, so the investigator ranked subjects by their subtotals for Approach, Plan, and Result: A_i was equal to the sum of the Approach scores for subject i across the six problems; P_i was equal to the sum of the Plan scores; and R_i was equal to the sum of the Result scores. Thus, subject j who achieved scores of (1, 1, 0), (1, 2, 2), (0, 0, 0), (1, 2, 1), (1, 1, 1), and (1, 1, 2) for his Approach, Plan, and Results, respectively, attained subscores of $A_j = 5$, $P_j = 7$, and $R_j = 6$.

Ranking B was based on A_i, P_i, and R_i, but gave priority to subjects who demonstrated an understanding of the most problems. By this system, the highest A_i score was ranked first. In case of ties, the subject with the highest P_i scores received the next rank. If subjects were still tied, then the highest R_i received the next rank. If ties existed for all three scores, the ranks were averaged.

Ranking C was similar to Ranking B, but it emphasized the subject's plans and processes. The P_i scores of subjects were used first to determine a ranking, and the A_i and R_i scores were compared in that order if ties occurred. Table 4 presents the A_i, P_i, and R_i scores, the total scores, and Rankings A, B, and C.

Table 4

Interview Test Scores and Rankings A, B, and C

Subject	Approach subtotal A_i	Plan subtotal P_i	Result subtotal R_i	Total interview test score	Ranking		
					A	B	C
1	5	5	4	14	8	6	9
2	2	3	4	9	14[a]	19.5[a]	19.5[a]
3	2	3	3	8	19[a]	21	21
4	5	7	6	18	5[a]	4.5[a]	4.5[a]
5	3	4	1	8	19[a]	15	12
6	1	1	1	3	28.5[a]	29	29
7	2	2	1	5	24.5[a]	24.5[a]	24.5[a]
8	2	3	4	9	14[a]	19.5[a]	19.5[a]
9	6	7	6	19	3	2	3
10	2	2	3	7	22	22	22
11	4	3	1	8	19[a]	11	16
12	5	3	1	9	14[a]	7	14
13	2	2	2	6	23	23	23
14	2	1	1	4	26.5[a]	26	27
15	6	10	8	24	1	1	1
16	1	2	1	4	26.5	28	26
17	3	4	3	10	10.5[a]	13.5[a]	10.5[a]
18	2	2	1	5	24.5	24.5[a]	24.5[a]
19	4	7	7	18	5[a]	8	6
20	3	3	2	8	19[a]	17	18
21	3	6	4	13	9	12	8
22	3	4	3	10	10.5[a]	13.5[a]	10.5[a]
23	3	3	3	9	14[a]	16	17
24	0	0	0	0	30.5[a]	30.5[a]	30.5
25	4	6	7	17	7	9	7
26	5	8	7	20	2	3	2
27	5	7	6	18	5[a]	4.5[a]	4.5[a]
28	2	1	0	3	28.5[a]	27	28
29	0	0	0	0	30.5[a]	30.5[a]	30.5[a]
30	2	4	2	8	19[a]	18	13
31	4	3	2	9	14[a]	10	15

Note. Subtotals were a subject's partial scores summed across the six interview problems.

[a]Ties occurred.

As can be seen in Table 4, Rankings A, B, and C agree on the ranks assigned to subjects 7, 10, 13, 15, 18, 14, and 29 and are similar in the other ranks. Since four pairs of subjects had identical subscores, Rankings B and C each produced four pairs of ties, and any other ranking system based on ordering A_i, P_i, and R_i would have had similar results.

Audio- vs. videotaping. The physical differences between audio- and videotaping are immediately apparent. Instead of a single tape recorder which the observer can operate alone, videotaping requires at least one camera, special lighting, and a technical assistant. To effectively capture a subject's actions and writing, more than one prefocused camera or a single camera which can be refocused is needed. Compared to audiotaping, the equipment and technical assistance necessary for videotaping is more costly to the investigator and perhaps more distracting to the subject.

In this study, the disadvantages of videotaping were offset by the variety of information which could be captured. Physical actions, nervous habits, and unspoken problem-solving procedures were noted on the videotape. For example, subjects reread the problem or parts of it silently, but the video record clearly indicated their behavior as they followed the sentences with their eyes or pencil, moved their lips, or asked a question immediately after staring at a problem. The 16 videotaped subjects produced 95 of these silent rereading behaviors, which would not have been evident on audiotape.

Another problem-solving strategy easily missed on audiotape occurred when subjects drew or modified a diagram without verbally indicating their actions. Problem 5 on the IT was solved by five audiotaped subjects with a sketch of a ladder, but the coder had to rely on completed diagrams and the subjects' verbalizations to speculate on modifications made during the solution attempts for the audiotaped subjects.

While the advantages of videotaping for recording subject behaviors were obvious, performance differences due to the videotaping procedure were possible. The investigator suspected that videotaped subjects spent less time solving the test problems and that their haste resulted in lower scores than the audiotaped subjects earned. These suspicions gave rise to hypotheses one (H1) and two (H2):

Hypothesis H1: The mean score on achievement for videotaped subjects equals the mean score on achievement for audiotaped subjects.

Table 5

Analysis of Variance for Total Interview Test Scores

Source	*df*	*MS*	*F*	*p*
Treatments	1	0.24	.006	1.00
Error	29	38.31		

Table 6

Analysis of Variance for Subjects' Total Solution Times on the Interview Test

Source	df[a]	MS	F	p
Treatments	1	101.00	3.97	.10
Error	27	25.44		

[a]Due to erasure of tape, two subjects' protocols could not be timed.

Hypothesis H2: The mean solution time of the videotaped subjects equals the mean solution time of the audiotaped subjects.

The ANOVA for H1 and H2 are reported in Tables 5 and 6, respectively.

As Table 5 indicates, null hypothesis H1 could not be rejected. The very low ratio of .006 was an indirect result of the close similarity of the videotaped and audiotaped subjects' total scores. The videotaped subjects mean score was 9.7 with a standard deviation of 5.8, while the audiotaped subjects achieved a mean score of 9.9 with a standard deviation of 6.2.

As shown in Table 6, the significance level of .05 was not reached and H2 could not be rejected. However, the *F* ratio of 3.97 was significant below the .10 level, and the analysis suggested there were some treatment differences. The videotaped subjects' mean solution time was 16.7 minutes and the audiotaped subjects' mean time was 13.0 minutes, contradicting the investigator's belief that the subjects performing in front of a camera may have worked more hastily.

Lucas (personal communication) suggested that coding videotaped protocols took less time than coding audiotaped protocols. His observation was tested with hypothesis H3.

Hypothesis H3: The mean coding time for audiotaped protocols equals the mean coding time for videotaped protocols.

The analysis of variance of these data is reported in Table 7.

Table 7

Analysis of Variance for Coding Times

Source	df[a]	MS	F	p <
Treatments	1	0.68	.002	1.00
Error	27	292.09		

[a]Due to erasure of tape, two coding times could not be measured.

Table 7 illustrates that the extremely low *F* ratio of .002 did not reach the .10 significance level. Thus, H3 was not rejected. The means of 42.3 (videotaping) and 42.6 (audiotaping) minutes of coding time per subject and

variances of 17.3 (videotaping) and 15.8 (audiotaping) indicated that coding time distributions were nearly identical. However, the videotaped protocols lasted 251 minutes and took 635 minutes to code, while the audiotaped protocols were 182 minutes long and took 597 minutes to code. Thus, 1 minute of audiotape took an average of 3.28 minutes to code, but 1 minute of videotape took only 2.53 minutes to code. Coding 1 minute of videotape took only 75% as long as coding 1 minute of audiotape, a savings of approximately 22%.

Statistical Analyses of Rankings

The feasibility of using a written instrument as a substitute for the interview and coding procedure depended upon the relationships between the data from the written tests and the interview tests. Two written tests, the WT and the WT2, were administered and three rankings, A, B, and C, were developed from the IT. The initial statistical findings are reported next, followed by an explanation of the exploratory procedures used to seek additional rankings.

Relationships of the written and interview tests. The rankings from the written and interview tests yielded two possible comparisons. A Pearson product-moment correlation coefficient r_{xy} (Hays, 1963, p. 497) was computed between the raw scores (number correct) on the written tests and the interview test total and subtotal scores. For each correlation coefficient, a hypothesis that the population statistic P_{xy} equals zero was tested by a t-test with N-2 degrees of freedom.

In addition to the correlation between scores, the relationship between the rankings developed from the tests was also measured. Kendall's τ (Hays, 1963, p. 642) with ties was computed for the association between the rankings, and the significance level of τ was found by computing z values. Because of ties within rankings, Goodman's and Kruskal's γ statistic (Harp, 1963, p. 655) was computed to provide a simpler interpretation of Kendall's τ. The correlations and rankings statistics are presented in Table 8.

Table 8

Correlation and Ranking Statistics for the Interview Test and the Written Tests

	r_{xy}	τ	$p(\tau)$	γ
WT and Ranking A	.61*	.44	.001	.48
WT and Ranking B	.40**	.33	.007	.34
WT and Ranking C	.59*	.39	.002	.41
WT2 and Ranking A	.64*	.49	.001	.52
WT2 and Ranking B	.48**	.38	.002	.40
WT2 and Ranking C	.61*	.45	.001	.46
(WT + WT2) and Ranking A	.68*	.50	.001	.52

*Significant at the .001 level in a two tailed t-test of H_0: $P_{xy} = 0$.

**Significant at the .05 level in a two tailed t-test of H_0: $P_{xy} = 0$.

As reported in Table 8, none of the correlation coefficients between the seven pairs of written and interview test scores attained the desired minimum of .71, although the combined scores of the WT and the WT2 produced an encouraging correlation coefficient of .68 with the total IT score. Two pairs of scores (WT and Ranking A; WT2 and Ranking C) each produced a correlation of .61. All seven correlation coefficients resulted in *t*-test values significant at the .05 level. Thus, the hypothesis that no correlation exists between written and interview test scores was rejected.

The associations between the rankings reported in Table 8 resulted in low but statistically significant values. Kendall's τ ranged from a low of .33 for WT and Ranking B to a high of .50 for (WT + WT2) and Ranking A. Kruskal's γ ranged from .34 for WT and Ranking B to .52 for WT2 and Ranking A and (WT + WT2) and Ranking A.

Exploratory Procedures

Latent partitioning and clustering were the statistical analyses used to find underlying patterns among subjects and to possibly produce other ranking schemes. Because the computer program for latent partitioning was not available, Guttman-Lingoes multidimensional scaling was substituted. A similarity measure D based on subscores for Approach, Plan, and Result was computed between each pair of the 31 subjects and was used in both analyses.

The Guttman-Lingoes multidimensional scaling program (Lingoes, 1973) searches for underlying patterns or structures among similarity measures. The program then represents the structure in a spatial model by assigning coordinates to the subjects and computes stress values to measure the agreement, the order of the spatial distances, and the order of the similarity measures. High agreement is indicated by low stress values. A second measure, the coefficient of alienation, deals with a type of monotonicity criterion for the relationship between distance and similarity measures. The coordinates, stress values, and coefficients of alienation for one, two, three, and four dimensions were produced by the Guttman-Lingoes program. The one-dimension results accommodated a ranking which closely paralleled Ranking A.

Johnson's (1967) max clustering algorithm was the second exploratory procedure used to group subjects according to a structure underlying the similarity measures. The program defines a sequence of partitions of a set of objects and uses similarity values to determine diameters of the subset. The max procedure constructs hierarchical partitions containing subsets of minimum diameter and assigns a partition rank to each pair of objects. Inspection of Johnson's clustering results revealed a pattern strongly resembling the ranking scheme developed from one-dimensional scaling.

Conclusion

This study attempted to find valid procedures for measuring students' mathematical problem-solving achievement. The commercial tests which were examined seemed inadequate to assess that achievement. Taped thinking aloud interviews and an associated coding system capture and classify mathematical problem-solving behaviors much better than commercial tests, but these complex interview procedures are not feasible for large scale use. A written test having high concurrent validity with experimental interview results would be a useful alternative. The first question this study examined was the feasibility of producing such a written test.

The physical and statistical qualities of the WT and WT2 indicated that they were suitable for administration to seventh-grade students in classrooms. Experimental Groups A and B required, on an average, less than 27 minutes to complete the WTs, and no great deviation would be expected when parallel forms of this test are used by other seventh-grade classes. The average Hoyt reliability (Hoyt, 1941) of both written tests across all groups was an acceptable .79.

The feasibility of the written test was measured by its prediction of seventh graders' problem-solving performance and ranks on the IT. The product-moment correlation coefficient was .61 between the IT and WT ranks and .64 between the IT and WT2 ranks. Though both values were highly significant ($p < .001$), neither the WT nor the WT2 attained the minimum correlation of .71. Thus, the written test was declared not presently feasible for predicting mathematics problem-solving performance as measured by the thinking aloud procedure and coding scheme. Future research could improve the test statistics by replicating this study with a larger population, using a longer written test, using more mathematical problems on the interview test, using a revised scoring system, or screening the WT items and IT problems to select only those which have high correlation to other items.

The second main question of the study was, "Is it possible to assess, separate, and rank seventh graders according to their problem-solving protocols?" The answer appears to be positive. A variation of Lucas's (1972) coding system was applied to verbal problem-solving protocols with a high degree of agreement and reliability. Rankings A, B, and C' were derived from the scores awarded by Lucas's point system and provided high rank order agreement measures. The order imposed by Ranking A was consistent with similarities and patterns detected among the subjects by scaling and clustering analyses. Statistics comparing rankings also indicated a high degree of agreement. Future research will be needed to refine the application of multidimensional scaling and clustering procedures to measures of mathematical problem-solving achievement.

Probably the most important finding of this study was the answer to the first research question. The question was "How well do the thinking aloud procedure and related coding scheme capture and classify the mathematical problem-solving behaviors of seventh graders?" The answer appears to be, "Not very well." The behavior of the students during the thinking aloud interviews raised critical questions about the reliability and validity of the information recorded. The seven subjects who were obviously anxious were unlikely to exhibit their normal problem-solving behaviors. An additional six subjects gave more subtle indications that they were anxious. Therefore, almost one third of the subjects were not performing normally. Other subjects who had difficulty talking while thinking add to the suspicion that the procedure did not adequately represent problem-solving behaviors and that it may not be highly valid or reliable with seventh graders. Systematic examination beginning with first graders and continuing through adults should detect general trends in ability to think aloud with increased mental maturity.

Videotaping has a distinct advantage over audiotaping because it can detect silent rereading, drawing and altering diagrams, and written computation. Videotaped protocols also take less time to code per minute of tape. Future investigators will need to decide if the extra information and time saved is worth the expense of videotaping. The author's dissertation contains a more complete account of this study (Zalewski, 1974).

Chapter 8

Development of a Test of Mathematical Problem-solving which Yields a Comprehension, Application, and Problem-solving Score

Diana C. Wearne

Mathematicians and mathematics educators agree on the importance of developing the problem-solving abilities of children. The Cambridge Conference on School Mathematics (Educational Services, Inc., 1963), the College Entrance Examination Board (1959), and the National Advisory Committee on Mathematics Education (NACOME) (1975), among others, all stressed the importance of problem solving in school mathematics programs.

Following these recommendations, problem solving has become prominent in text series. One such series, *Developing Mathematical Processes* (DMP) (Romberg, Harvey, Moser, & Montgomery, 1974, 1975, 1976), views problem solving as the vehicle for achieving its program goals (Romberg & Harvey, 1969). DMP is a research based, individually guided instructional program in elementary mathematics developed by the staff of the Analysis of Mathematical Instruction Project at the Wisconsin Research and Development Center for Cognitive Learning at the University of Wisconsin. The authors refer to the program's activity approach to learning as learning through problem solving.

There has been some disagreement on the type of problems to include in a mathematics program. Kline (1973) and others have consistently and broadly criticized the application problems in mathematics texts as having little in common with real life situations. Nelson and Kirkpatrick (1975) also have emphasized real life situations. Others believe the real life category to be too restrictive and have advocated any problem available for mathematics analysis (NACOME, 1975). However, there is no disagreement on the importance of including problem-solving activities in mathematics programs.

Polya's (1962) frequently quoted statement on the importance of problem solving voices the feelings of virtually all mathematics educators:

> What is know-how in mathematics? The ability to solve problems — not merely routine problems but problems requiring some degree of independence, judgment, originality, and creativity. (p. viii)

In addition to advocating the inclusion of problem-solving material in school mathematics programs, mathematics educators are engaged in research on problem-solving behavior, particularly the heuristics of problem solving. It appears that a measure of problem-solving ability is needed to determine how well problem-solving abilities are being developed. The instrument described in this chapter was developed in response to this need.

Background of the problem

An individually administered test of problem-solving behavior not only produces a score but also provides an opportunity to observe the child solving the problem. The child can be asked how the problem was solved, or if the child was unsuccessful, what path of reasoning was followed and what aspects of the problem were confusing.

However, the limited amount of time usually allowed for assessing problem-solving behavior makes a group-administered test necessary. In group-administered testing, however, the examiner is unable to identify which subjects were unable to solve the problem because they did not understand the information presented, had not mastered the concepts or processes needed, or could not apply the prequisite concepts or processes even though they knew them.

A cursory examination of existing group-administered tests of problem-solving behaviors reveal that the authors of these tests apparently have defined problem solving in terms of verbal, mainly one-step, problems. The operation required to solve the problem is frequently implied by the wording of the problem itself; for example, asking "What is the area of . . . ?" or "How much more . . . ?"

The Stanford Achievement Test (Kelley, Madden, Gardner, & Rudman, 1964) contains a section entitled "Applications." However, the *Directions for Administering, Intermediate I Battery* refers to that portion of the tests as measuring problem solving. Examples from the section include:

2. Don is delivering papers to earn more money. He had 150 papers to deliver an hour ago. He has delivered 90 of them now. How many are left to deliver?

25. If two pencils cost 15¢, how many can you buy for 30¢?

Intermediate I Battery is designed for children in grade 4 and the first half of grade 5.

The Modern Math Understanding Test, Form C, Multilevel Edition (Science Research Associates, 1966) classifies 12 items as being problem solving or application problems. All of the items may be described as simple applications or concept assessments. Examples from this test are as follows:

9. On a certain map a distance of 1 inch represents 200 miles. If the distance between two towns is 2½ inches on the map, how many miles apart are the towns?

17. Which numeral must be placed in the box to make the following sentence true?

$6 \times 2 = (\square \times 2) + (1 \times 2)$

35. The perimeter of this rectangle is _____ .

(A pictured rectangle is shown with the measurements 11 inches and 4 inches on adjoining sides.)

Other items in this set refer to the concepts of relatively prime numbers and equivalent fractions and to adding the lengths of line segments together.

Other tests such as the *Iowa Test of Basic Skills* (Hieronymus & Lindquist, 1971) and the *Metropolitan Achievement Tests* (Durost, Bixler, Wrightstone, Prescott, & Balow, 1971a, 1971b) also contain problem sets identified as assessing problem-solving behaviors. The comments made about the *Stanford Achievement Test* and the *Modern Math Understanding Test* apply to these tests as well.

An alternative to the standardized tests is a test consisting of items which conform to the investigators' definition of problem-solving (Kilpatrick, 1967; Zalewski, 1974); Zalewski's study is reported in Chapter 7.

Studies have reported the primary factors related to success in problem-solving are reading to note details, understanding of the vocabulary, mastery of the necessary computation skills, and knowledge of the relevant mathematical concepts (Chase, 1960; Johnson, 1944). Alexander (1960) and Treacy (1944) reported that good and poor problem solvers differed in aspects of reading. Specific instruction in quantitative vocabulary was found by Vanderlene (1964) to increase problem-solving scores. Bogolyubev (1972) noted that children could misconstrue words in verbal problems, thus misunderstanding the problem to be solved. In another study, Egan and Green (1973) reported that individual differences in prerequisite knowledge were more important for "discovery" learning and creative problem-solving than for "rule" learning. Norman (1950) found that merely possessing a necessary computational skill did not imply a child would be able to solve verbal problems using that skill.

The research reported above underscores the possibility that lack of familiarity with the vocabulary in a question or not possessing an appropriate level of concept attainment may result in an incorrect response to a problem-solving question.

The Test

When presented with a problem-solving task, ideally the child should understand the information in the task description and have the prerequisite skills for solving the problem. Without this understanding and skill, there is no certainty that a child cannot solve a problem simply because he or she has responded incorrectly. For example, if the task involves the concept of area and the child is unfamiliar with this concept, a wrong answer will not necessarily indicate inability to solve the problem, but merely that the child is unfamiliar with the underlying concept.

The test described in this chapter sought to evaluate the child's understanding of the vocabulary and mastery of the prerequisite concepts or processes of each problem-solving question. Such a test could yield more information about the child and provide a "truer" measure of problem-solving ability by considering, as a measure of that ability, only those problems for which the child has mastered the prerequisites.

The test was to produce three scores: a comprehension score, an application score, and a problem-solving score. To accomplish this, the test contained clusters of items called superitems; each superitem consisted of a comprehension item, an application item, and a problem-solving item.

The comprehension item assessed the child's understanding of the implicit or explicit information in the item stem. The application item assessed the child's mastery of a prerequisite concept or skill of the problem-solving item by applying it in a fairly straightforward way. The third item was the problem-solving one.

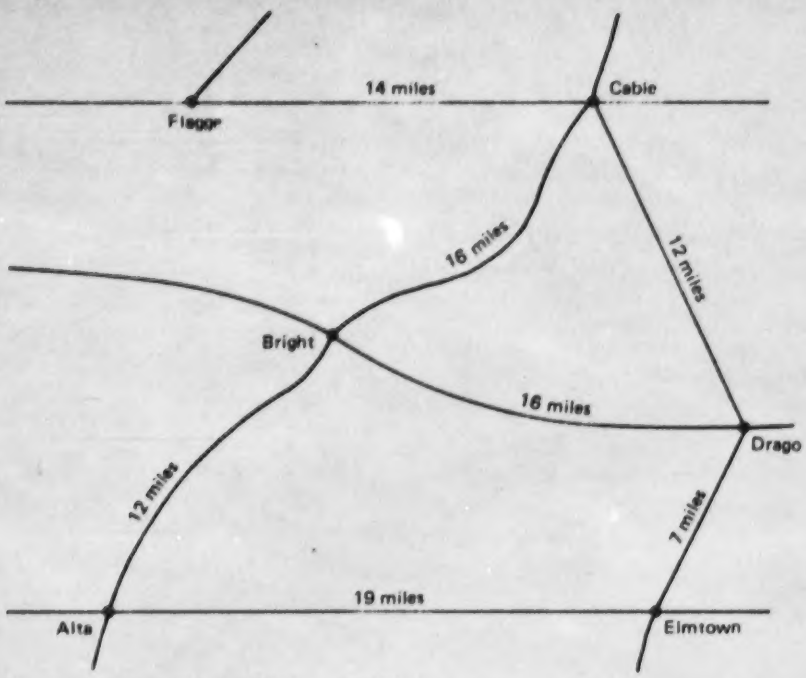
A problem situation was defined as a situation which posed a question whose solution was not immediately available; that is, a situation which did not lend itself to immediate application of some rule or algorithm. An effort was made to construct short items that did not appear impossible for the child to solve. A guide for the construction might be found in Hilbert's (1906) comment.

> A mathematical problem should be difficult in order to entice us, yet not completely inaccessible, lest it mock at our efforts. It should be to us a guidepost on the hazy paths to hidden truths and ultimately a reminder of our pleasure in the successful solution. (p. 59)

Although the application and problem-solving items referred to a common unit of information, the item stem, the items were independent to the extent that the response to the application item was not needed to answer the problem-solving item.

Examples of Superitems

The first example of a superitem appears in Figure 1. The first of the three questions comprising the superitem is the comprehension item. This



The distance from Alta to Bright is: 7 miles

12 miles

16 miles

19 miles

The shortest distance from Alta to Drago is: through Bright

through Cable

through Elmtown

through Flagge

The sign

BRIGHT	16
ELMTOWN	19

should be placed: in Drago

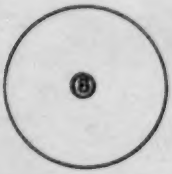
in Alta

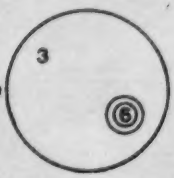
in Flagge

in Cable

Figure 1. Example 1 of superitems.

In Circleland, people write when they mean 475 and they write when they mean 81.

In Circleland, people write  when they mean 8. TRUE FALSE

What do they mean when they write  ?

36

63

630

603

306

What do they mean when they write ?

4,526

40,526

4,562

45,620

45,260

Figure 2. Example 2 of superitems.

seeks to determine if the child understands the information on the map; specifically, that the numbers on the map are distances in miles between towns.

The second item is the application item. To respond correctly, the child must be able to read the map, identify the distances referred to by the question, and correctly add these distances together.

To respond correctly to the problem-solving item, the final one in this superitem, the child must be able to read the map, add two distances, and find a position on the map corresponding to the given distances.

The second example of a superitem is shown in Figure 2. Once again, the first item is the comprehension item. This assesses the child's understanding of the information presented in the item stem. The application item is a direct application of the information in the item stem. The problem-solving item asks the child to arrive at a generalization of the information in the item stem.

Development of the Test

A further impetus to developing a test of problem solving was the desire of the Analysis of Mathematics Instruction (AMI) project staff to include a problem-solving measure as one component of the terminal accountability tests for *Developing Mathematical Processes* (DMP) (Romberg et al., 1974, 1975, 1976).

Because the test described here was to be a model for the problem-solving portion of the DMP terminal accountability tests, constraints were imposed on the application and problem-solving questions. The application items had to measure mastery behaviors of the program and the problem-solving items had to measure behaviors beyond the mastery level of children at the end of the fourth grade; the test described in this chapter was designed for fourth-grade children. For example, in DMP children are expected to master finding the area of a rectangle or a figure composed of rectangular regions by the end of the fourth grade; children are not expected to master finding the area of a nonrectangular figure or a figure not composed of rectangles. Thus, finding the area of a rectangle would be an appropriate application item and finding the area of a nonrectangular region would constitute a problem-solving question.

Question Raised by the Format of the Test

A test composed of superitems yields more information than one containing only problem-solving questions; however, structuring the test in this manner raises some questions.

The questions are as follows:

1. Does asking a series of questions have a facilitating or debilitating effect on the response to the questions? In particular, does the inclusion of a

Table 1

Description of the Tests

Test	Type of items		
	Comprehension	Application	Problem-solving
C	x		
A		x	
P			x
CA	x	x	
CP	x		x
AP		x	x
CAP	x	x	x

comprehension and an application item have an effect on the response to the problem-solving item?

2. How should the reliability of the test be estimated?

3. What type of validity will be obtained?

4. To what extent is the model for the test supported by the test results? That is, are correct responses to the comprehension and application items required for a correct response to the problem-solving question?

A discussion of the procedures followed in investigating these questions is contained in the next section.

The Results

The Effect of the Superitem Format on Item Response

The investigation focused on the effect of the comprehension, application, and problem-solving items upon each other; that is, the effect of asking multiple questions on the same unit of information upon the response to those questions. The inclusion of the comprehension and application items provides more information than a test containing only the problem-solving items. However, if including the additional items affects the response to the problem-solving items, then the strength of the problem-solving test is diminished.

To determine the effect of the items upon each other, six tests consisting of subsets of the original set of items were constructed. Three of the tests contained only one of the three types of items; the remaining three tests contained two of the three types of items. Thus, there were a total of seven tests, the six tests containing subsets of the items and the complete test. The content of the tests is described in Table 1.

Each of the subset tests was administered to approximately 50 children. To minimize teacher effect as much as possible, each test was adminis-

Table 2

Number of Children Taking Each of the Tests

	Test							
School	C	A	P	CA	CP	AP	CAP	Total
1	10	14			13	11		48
2		12	12	14		11		49
3		8	7	11	9			35
4	19	13	25	10	14	12		93
5	9	3	5	7	10	12		46
6							90	90
7							53	53
8							68	68
9							37	37
10							31	31
11							38	38
Total	38	50	49	42	46	46	317	588

tered in four classes. Each of the four classes was in a different school except for two classes in the same school who took the C test. The children in each class were randomly assigned to one of two instruments. The complete test was administered either by the investigator or by a person who had been trained in responding to the children's questions; the other six tests were administered by the investigator. The number of children taking each test is given in Table 2.

The population for this study consisted of fourth-grade children from Wisconsin who had been using DMP (Romberg et al., 1974, 1975, 1976) for at least 1½ years. Those who took the complete test (CAP) either attended one of four schools in a city of 40,000 which is a suburb of a large city or they attended one of two schools in a suburb of a medium-sized city. The children taking the subtests attended schools in the following locations: a city of population 50,000, a small town, a suburb of a medium-sized city, and a medium-sized city.

It was difficult to determine an administration time for the six subtests. The administration time of a complete test is 45 minutes. However, it was not possible to merely apportion time based upon the number of items in a subtest. Two assumptions affected the administration time:

1. It was assumed the problem-solving items require more response time than the application items which, in turn, require more response time than the comprehension items.

Table 3

Test Administration Time

Test	Administration time (in minutes)
C	20
A	25
P	30
CA	35
CP	35
AP	35
CAP	45

Table 4

The Means, Variances, and Reliability Estimates of the Scales on Each Test Containing the Scale

Scale	Test	Number	Mean	Variance	Reliability
C	C	38	15.95	6.97	.49
	CA	42	17.02	7.54	.58
	CP	46	15.93	10.06	.65
	CAP	317	15.53	9.74	.63
A	A	50	11.08	17.22	.78
	CA	42	10.50	15.43	.76
	AP	46	8.61	15.09	.75
	CAP	317	10.01	13.15	.71
P	P	49	4.24	11.02	.74
	CP	46	2.93	4.37	.49
	AP	46	3.46	7.81	.69
	CAP	317	3.30	6.13	.60
CA	CA	42	27.52	38.74	.82
	CAP	317	25.54	37.41	.80
CP	CP	46	18.87	18.96	.69
	CAP	317	18.82	22.53	.73
AP	AP	46	12.07	38.86	.84
	CAP	317	13.31	31.26	.80
CAP	CAP	317	28.84	62.18	.84

Table 5

ANOVA Summary Table for Scores

Scores	*SS*	*df*	*MS*	*F*
Comprehension Scores				
Between groups	86.98	3	28.99	3.11*
Within groups	4098.70	439	9.33	
Application Scores				
Between groups	157.42	3	52.47	3.75*
Within groups	6309.09	451	13.99	
Problem-solving Scores				
Between groups	48.25	3	16.08	2.42
Within groups	3015.51	454	6.64	

*$p < .05$.

2. Since several items frequently shared the same item stem, it was assumed that the child did not have to process two pieces of information and, hence, would not need the sum of the times needed to respond to the items individually.

Using these assumptions, the administration times chosen were as follows: 20 minutes for the C test, 25 minutes for the A test, 30 minutes for the P test, and 35 minutes for the CA, CP, and AP tests. The administration times are summarized in Table 3.

As noted previously, the test contains three types of items; for the remainder of this paper, these categories of items will be referred to as *scales*. The number of subjects, means, variances, and reliability estimates for the Comprehension, Application, and Problem-solving scales on each of the tests containing them are reported in Table 4.

The means of the scales on the various tests were compared to determine if the items had an effect on one another. Consequently, it was more serious to neglect to identify a significant difference than to identify more significant differences than actually exist. Stated another way, a Type II error was of greater consequence than a Type I error. To avoid a Type II error a more generous alpha level was chosen than would be used if one were primarily interested in avoiding a Type I error. The results of the analysis of variance for each of the three scales are summarized in Table 5.

An *a posteriori* multiple comparison test was used to determine if the difference between the means on the same scale were significantly different on the tests containing the scale. Normally one would use a planned comparison procedure rather than a post hoc procedure when it is essential that a Type II error not be committed. However, all pairs of means were to be used, and this use violates the independence required by a planned comparison test. Independence is not required by post hoc procedures. The procedure used in this study was Scheffe's (1953) method of comparing all possible means.

The probability of overlooking a true difference from zero is greater in the post hoc procedures than in the planned method. Therefore, an alpha level of .10 was selected. The probability of obtaining at least one spuriously significant comparison using a post hoc procedure equals alpha (Hays, 1973).

Two differences were significant at the .10 level, both of which were also significant at the .05 level. One of the significant differences was found among the comprehension scores and one among the application scores; no significant differences were found among the problem-solving scores. The significant difference among the comprehension scores was between the mean comprehension score on the CA test and the mean comprehension score on the CAP test. The significant difference among the application scores was between the mean application score on the A and the AP tests.

Two hypotheses may be advanced to account for these differences. One hypothesis is that the items have an effect, under certain conditions, upon one another; the other is that the children did not have the same amount of time to respond to a particular group of items on all the tests containing that group. These hypotheses will be examined in turn.

The higher mean comprehension score on the CA test may have been due to a facilitating effect of the application items. This effect could occur if the child responded to the comprehension item after responding to the application item, a retroactive effect, or if the application items had a stimulating effect upon the response to the comprehension items. The CAP test also contained the application items but it could be that the facilitating effect of the application items on this test was counterbalanced by a debilitating effect of the problem-solving items or that a retroactive effect does not take place in the presence of the problem-solving items. However, if a debilitating effect was produced by the problem-solving items, this effect should also have appeared on the CP test, which was not the case. A possible argument still remains that the debilitating effect of the problem-solving items upon the comprehension items only takes place in the presence of the application items; such an intricate dependency is possible, but unlikely.

The second hypothesis concerning the administration times of the tests offers another possible explanation for the significant difference. The children taking the CA test may have had more time to respond to the comprehension items than the children taking the other three tests containing the comprehension items. The CA and CP tests both had an administration time of 35 minutes; however, the application items are assumed to be less difficult than the problem-solving items. Thus, the children may well have had more time to respond to the comprehension items on the CA test than on the CP test. Although the mean comprehension score on the CA test was not significantly different from the mean comprehension score on the CP test, the mean on the CP test differed from the mean comprehension score on the CAP test by .40 points. The administration time for the CAP test was 45 minutes, 10 minutes

longer than the administration time of the CA test, but the CAP test included the problem-solving items. The problem-solving items are assumed to be more difficult and, hence, require more response time than the other two types of items. Therefore, of the two possible explanations for the significant differences among mean comprehension scores proposed in the preceding paragraphs, the more reasonable appears to concern the administration time of the tests.

The significant difference among mean application scores was between the score on the A test and on the AP test. One possible explanation for this significant difference is the problem-solving items on the AP test had a debilitating effect upon the response to the application items. However, this same effect did not occur on the CAP test though it can be argued that the effect of the problem-solving items on the CAP was tempered by the effect of the comprehension items on the application items. This interdependency may be in effect but does not appear likely.

Once again, another explanation for the difference lies in the administration time allotted for the tests. The children had 25 minutes to respond to the application items on the A test but only 35 minutes to respond to both the application and the problem-solving items on the AP test. The CAP test contained the comprehension items in addition to all the items contained on the AP test. Although the comprehension items require the least response time, the administration time of the CAP test was 10 minutes longer than that of the AP test. This provided more time for the children to respond to the application and problem-solving items than they had on the AP test. The difference between the mean application scores on the A and CAP tests was not significant.

Thus, of the two hypotheses advanced to explain the significant difference existing between the mean application score on the A and AP tests, the more reasonable appears to be the effect of the time allotted to respond to the items on the test.

Conditional Probabilities Associated with the Items

The superitem model assumes that the comprehension and application items assess prerequisite behaviors for the problem-solving item. The data for determining how well the superitems fit the model are presented in this section.

In the model, a correct response to the comprehension item was a prerequisite to responding correctly to the application item, which in turn was a prerequisite to answering the problem-solving item. To determine to what extent this was true, the following conditional probabilities were computed:

Prob (Comprehension item correct | Application item correct)

Prob (Comprehension item correct | Problem-solving item correct)

Prob (Application item correct | Problem-solving item correct)

Table 6

Conditional Probabilities Associated with the Superitems

Item	P(c | a)[1]	P(c | p)[2]	P(a | p)[3]	P(c ∩ a | p)[4]
1	.98	.99	.83	.83
2	.98	.99	.92	.91
3	.90	.85	.85	.74
4	.78	.80	.66	.63
5	.95	.97	.94	.92
6	.81	.74	.83	.66
7	.79	.77	.90	.71
8	.96	1.00	.80	.80
9	.91	.91	.83	.78
10	.70	.76	.47	.41
11	.78	.71	.76	.62
12	.85	.77	.62	.52
13	.63	.77	0	0
14	.93	.97	.80	.77
15	1.00	.90	.29	.29
16	.72	.59	.24	.24
17	1.00	1.00	.95	.95
18	.91	.97	.83	.79
19	.72	.52	.13	.11
20	.94	.86	.64	.64
21	.75	.74	.67	.50
22	.98	.94	.86	.86

[1]Conditional probabilities: Prob (comprehension item correct | application item correct).

[2]Prob (comprehension item correct | problem-solving item correct).

[3]Prob (application item correct | problem-solving item corrrect).

[4]Prob (comprehension and application items correct | problem-solving item correct).

Prob (Comprehension and Application items correct | Problem-solving item correct)

The values for the Prob (Comprehension item correct | Application item correct) ranged from .63 to 1.00 with a mean conditional probability of .86. The Prob (Comprehension item correct | Problem-solving item correct) varied from .52 to 1.00 with a mean probability of .82. The mean conditional probability of Prob (Application item correct | Problem-solving item correct) was .67; the probabilities ranged from 0 to .95. The Prob (Comprehension and Application items correct | Problem-solving items correct) ranged from 0 to .95 with a mean probability of .62. The conditional probabilities for the items are listed in Table 6.

A partial ordering of the superitems based upon their conditional probabilities for each of the four conditional probabilities of interest is contained in Table 7.

Table 7
Ordering of the Conditional Probabilities of the Superitems

Conditional probability	P(c | a)	P(c | p)	P(a | p)	P(c ∩ a | p)
	Superitem numbers			
.90 - 1.00	1,2,3,5,8, 9,14,15,17, 18,20,22	1,2,5,8,9, 14,15,17, 18,22	2,5,7	2,5,17
.80 - .89	6,12	3,4,20	1,3,6,8,9, 14,18,22	1,8,22
.70 - .79	4,7,10,11, 16,19,21	6,7,10,11, 12,13,21	11	3,7,9,14,18
.60 - .69	13		4,12,20,21	4,6,11,20
.50 - .59		16,19		12,21
.40 - .49			10	10
.30 - .39				
.20 - .29			15,16	15,16
.10 - .19			19	19
.00 - .09			13	13

For a superitem to agree with the model, all four of the conditional probabilities should reflect that agreement. Superitems were divided into three categories based upon their conditional probabilities; the divisions were arbitrarily selected by the investigator. If .75 is a minimum conditional probability at which to consider a superitem acceptable, then 10 of the superitems satisfied this criterion; that is, 10 of the superitems had four conditional probabilities of at least .75. There were seven additional superitems all of whose conditional probabilities were less than .75 but which were at least .50; these superitems were considered marginally acceptable. Five superitems did not have all four conditional probabilities of at least .50. Table 8 lists the numbers of the superitems categorized as Acceptable, Marginally Acceptable, and Unacceptable.

There were five superitems whose conditional probabilities placed them in the Unacceptable category. They were superitems 10, 13, 15, 16, and 19. Four of these superitems contained problem-solving items which ranked among the six most difficult problem-solving items on the test. The difficulty levels of the problem-solving items for superitems 10, 13, 15, 16, and 19 were .05, .04, .07, .05, and .26, respectively; the mean difficulty level of all the problem-solving items was .15. These five superitems also contained application items which were among the six most difficult on the test. The difficulty levels of the application items were .43, .05, .08, .09, and .16 for superitems 10, 13, 15, 16, and 19, respectively; the mean difficulty level of all the application items on the test was .46. Therefore, for four of the five superitems, at most 7% of the children responded correctly to the problem-solving portion and for four of the five superitems, at most 16% of the children responded correctly to the application portion. These difficulty values may have contributed to the low

Table 8

Categorization of the Items on the Basis of Their Conditional Probabilities

Category	Conditional probability	Superitems with all four of the conditional probabilities at that level
Acceptable	.75 - 1.00	1, 2, 3, 5, 8, 9, 14, 17, 18, 22
Marginally acceptable	.50 - .74	4, 6, 7, 11, 12, 20, 21
Unacceptable	.00 - .49	10, 13, 15, 16, 19

conditional probabilities in that the probabilities were based on very small samples. For example, only 13 children responded correctly to the problem-solving item of superitem 13 from a sample of 317 children.

An examination of the five superitems has led the author to conclude that superitems 13 and 16 may have been placed in the Unacceptable category for reasons other than failure to fit the model. (For a detailed discussion of the superitems, see Wearne, 1976.)

Validity

Due to lack of established criteria against which the tasks could be validated, the only indicant of validity available was content validity as judged by a panel of experts.

A test may be said to have content validity if it measures something which a group of authorities asserts it does measure. The American Psychological Association and the American Education Research Association in their joint publication *Standards for Educational and Psychological Tests and Manuals* defines content validity as follows:

> Content validity is demonstrated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn (American Psychological Association, 1966, p. 12).

Thus, evaluating the content validity of a test is tantamount to evaluating the adequacy of a definition.

A panel of six judges was selected on the basis of their familiarity with the DMP materials and interest in problem-solving research. A constraint on developing superitems for the test had been suitability for fourth-grade children in the DMP program. Thus, the judges had to be familiar with the content of the DMP program to judge whether a required behavior represented routine application of a concept or algorithm (application item) or if the behavior represented a nonroutine application (problem-solving item).

The judges were given the definitions of the three categories of items (comprehension, application, and problem-solving) and the items from the test in a random order. They were asked to classify the items as comprehension items, application items, or problem-solving items. The definitions given the judges were as follows:

1. Comprehension Item: This item assesses the child's understanding of the information contained either implicitly or explicitly in the item stem. When the item is assessing information contained implicitly in the item stem, it may be thought of as assessing the understanding of the definition of a basic concept underlying the situation.

2. Application Item: This item is a fairly straightforward application of some rule or concept to a situation. This item may be thought of as assessing what is considered to be a mastery behavior in the DMP program at the end of the fourth grade.

3. Problem-solving Item: A problem situation is defined to be a situation which poses a question whose solution is not immediately available, that is, a situation which does not lend itself to an immediate application of some rule or algorithm. This item may be thought of as assessing behavior beyond the mastery level of DMP at the end of the fourth grade.

The judges' classification of the items was compared to the classification of the items on the test. The mean proportion of judges agreeing with the test classification was .84. The mean proportion of agreement with the test classification of comprehension items was .89, the mean agreement on the application items was .78, and the mean proportion of judges agreeing with the problem-solving classification was .84. Table 9 lists the proportion of judges agreeing with the test classification for each item.

There were nine items, out of a total of 66, on which fewer than two-thirds of the judges agreed with the test classification of the items; half of the judges agreed with the test classification on five of these nine items. One of the nine items was a comprehension item, five were application items, and three were problem-solving items. The tendency, when disagreeing with the application category, was to rate the item as problem-solving. Comprehension and problem-solving items were rated as application items when disagreeing with the test classifications for these items.

A measure of association was computed between the judges' classification of the items and the classification of the items used when developing the test. The strength of the association was computed to be .77; the index used was Cramer's Statistic ϕ' (Hays, 1973, p. 745). The number of items placed into each of the categories by the judges is shown in Table 10.

Of the six judges classifying the items, one judge's classification differed from the test classification on six of the 66 items. Three other judges differed on

Table 9

Classification of Items by Judges

Super-item	Test classification	Judges' classification			Proportion of judges agreeing with test classification
		C	A	P	
1	C	6	0	0	1.00
	A	5	1	0	.17
	P	0	5	1	.17
2	C	5	1	0	.83
	A	0	5	1	.83
	P	0	4	2	.33
3	C	5	0	1	.83
	A	2	4	0	.67
	P	0	0	6	1.00
4	C	6	0	0	1.00
	A	1	3	2	.50
	P	0	0	6	1.00
5	C	5	1	0	.83
	A	0	4	2	.67
	P	0	0	6	1.00
6	C	5	1	0	.83
	A	0	6	0	1.00
	P	0	2	4	.67
7	C	6	0	0	1.00
	A	0	4	2	.67
	P	0	0	6	1.00
8	C	5	0	1	.83
	A	0	6	0	1.00
	P	0	1	5	.83
9	C	6	0	0	1.00
	A	0	6	0	1.00
	P	0	0	6	1.00
10	C	6	0	0	1.00
	A	0	6	0	1.00
	P	0	0	6	1.00
11	C	6	0	0	1.00
	A	0	6	0	1.00
	P	0	0	6	1.00
12	C	6	0	0	1.00
	A	0	6	0	1.00
	P	0	0	6	1.00
13	C	6	0	0	1.00
	A	0	6	0	1.00
	P	0	2	4	.67
14	C	5	1	0	.83
	A	0	6	0	1.00
	P	0	2	4	.67

	C	6	0	0	1.00
15	A	0	3	3	.50
	P	0	0	6	1.00
	C	4	2	0	.67
16	A	0	2	4	.33
	P	0	0	6	1.00
	C	4	2	0	.67
17	A	0	3	3	.50
	P	0	0	6	1.00
	C	6	0	0	1.00
18	A	0	6	0	1.00
	P	0	0	6	1.00
	C	6	0	0	1.00
19	A	1	4	1	.67
	P	0	2	4	.67
	C	6	0	0	1.00
20	A	1	5	0	.83
	P	0	0	6	1.00
	C	5	1	0	.83
21	A	0	6	0	1.00
	P	0	0	6	1.00
	C	3	3	0	.50
22	A	1	5	0	.83
	P	0	3	3	.50

Table 10

Total Number of Items Placed into Each of the Categories by the Judges

Test classification	Judges' classification			Total
	C	A	P	
C	118	12	2	132
A	11	103	18	132
P	0	21	11	132

eight of the items. One judge differed 16 times and the remaining judge differed from the test classification on 18 occasions. Thus, four judges differed a total of 30 times and the remaining two judges differed a total of 34 times. The four judges who differed the fewest number of times had all been writers of the DMP program for a minimum of 2 years. Of the two judges who differed the greatest number of times, one had been a writer for only 1 year and had been involved with the revision of the K-2 materials rather than the 3-4 materials; the test contains more material related to the 3-4 component than the K-2 component. The judge who differed 18 times with the test's classifications had been associated with the DMP program since its inception but not as a writer of curriculum materials. It is quite possible that the familiarity gained by writing the materials may have made those four judges more aware of what constitues mastery behavior of the program than the two judges who were not involved with writing the 3-4 materials. Classification of the application and problem-solving items rests upon mastery objectives of the program.

No one judge consistently rated items as being in a higher or lower category than the classification of the items on the test. However, the two comprehension items rated as representing a problem-solving situation were rated by the same judge.

Reliability

Two estimates of reliability of the test were computed, the Hoyt reliability and KR-20 reliability. The Hoyt reliability was computed under the assumption that the test items were independent. However, if an indication of dependency was found, then the Hoyt estimate would be rendered unreliable. To satisfy this contingency, a generalized KR-20 was also computed for the test. This procedure was suggested by Cureton (1965) when discussing the problems associated with computing reliabilities of a test consisting of superitems.

$$r = 1 - \frac{MS_{\text{persons} \times \text{items}}}{MS_{\text{persons}}} \qquad \text{(Hoyt, 1941)}$$

$$r = \frac{n}{n-1}\left[1 - \frac{\sum \sigma^2_i}{\sigma^2_{\text{test}}}\right] \qquad \text{(Cureton, 1965)}$$

where n is the number of superitems, σ^2_i is the score-variance on each of the superitems, and σ^2_{test} is the variance of the scores on all the superitems.

Hoyt reliability estimates were also computed for each of the six subset tests as well as for each of the three scales, Comprehension, Application, and Problem-solving, contained in the six tests. Thus, a reliability estimate of the comprehension scores was obtained on the C, CA, CP, and CAP tests. The reliabilities for the Comprehension scale ranged from .49 on the C test to .65 on the CP test. The reliability for the Application scale ranged from .71 on the CAP test to .78 on the A test. The reliability estimates for the Problem-solving scale ranged from .49 on the CP test to .74 on the P test.

The reliabilities computed for the CA, CP, and AP tests were .82, .69, and .84, respectively. The reliabilities of these scales on the CAP were .80, .73, and .80, respectively. The reliabilities are given in Table 4.

The question of independence of the items was discussed in the section *The effect of the superitem format on item response*. The conclusion reached in this section was that the items did not have any effect upon one another and those effects that were noted were artifacts of the administration time. Thus, it may be assumed that the reliability estimate of .84 as reported by the Hoyt, which requires independent items, does not represent an inflated estimate of reliability.

The difference between the more conservative reliability estimate of .79 obtained by using a generalized KR-20 and the Hoyt estimate of .84 may exist because when items were placed into groups of three, as they were to form the superitems, variance among the items which had been part of the item-covariance in the Hoyt estimate constitute part of the item variance in the KR-20. Consequently, the KR-20 reapportions the total variance of the test with a greater proportion of the variance now being part of the item variance rather than the covariance of the items. This results in the ratio of item variance to total variance being larger and, hence, a lower reliability estimate.

Reliabilities were also computed for each scale, Comprehension, Application, and Problem-solving, on each test containing the scale; estimates of reliability were obtained for the entire CA, CP, and AP tests as well.

The reliabilities for the Comprehension scale on the four tests containing the scale varied from .49 on the C test to .65 on the CP test. The associated variances ranged from 6.97 on the C test to 10.06 on the CP test. The low reliability estimates are to be expected when taking the variances into account.

The reliability estimates associated with the Application scale ranged from .71 on the CAP test to .78 on the A test; the variances from 13.15 on the CAP test to 17.22 on the A test.

The range of reliability estimates reported for the Problem-solving scale was from .49 on the CP test to .74 on the P test; the associated variances ranged from 4.37 on the CP test to 11.02 on the P test. The reliability estimates are a reflection of the skewed distribution of scores on this scale. The range of the problem-solving scores on the CAP tests was from 0 to 15 with a mean of 3.30 and a standard deviation of 2.48. The mean proportion of children responding correctly to the 22 problem-solving items on the CAP test was .15.

The reliability estimates for the CA, CP, and AP tests were .82, .69, and .84, respectively. The reliabilities of these scales on the CAP test were .80, .73, and .80 for the CA, CP, and AP tests, respectively.

One way of increasing the reliability estimates of the Comprehension and Problem-solving scales would be to increase the variability in the responses to the items. However, to do so would violate the definition of these categories in both cases. The purpose of the comprehension item was to assess the child's understanding of the information contained, implicitly or explicitly, in the item stem. One would expect a relatively high degree of success on this item. The range of difficulty levels associated with the comprehension items was .35 to .98; the mean item difficulty was .71. An example of a comprehension item to which virtually all children responded correctly is in superitem 2; 96% of the children responded correctly to this item. The item (see Figure 1) asked the child to determine the number of miles between two towns, a distance given on the map. As stated earlier, the purpose of this comprehension item is to determine if the child understands that the numbers on the map represent distances, in miles, between towns. Therefore, this item is an appropriate comprehension item. To assess a more complex behavior would exceed the definition of the item type.

A problem situation was defined as one which poses a question whose solution is not immediately available, that is, a situation which does not lend itself to immediate application of some rule or algorithm. This definition permits a wide latitude in choosing problem situations since the only constraint is that the solution behavior required is not a mastery behavior from the first 65

topics of DMP and, at the same time, is a problem for which the children have the prerequisite conceptual and computational skills.

Traditionally, investigators of problem-solving behavior have not been particularly concerned with the reliability of their tests. The investigators apparently consider the task itself to be more important than any associate index of consistency. However, when administering a test to a group of children, an investigator would like to be reasonably confident that the results are reliable; that is, if the children took the test again, the same ordering would occur.

Reliability of the Problem-solving scale would increase if the items were less difficult. However, raising the reliability by deleting challenging questions would be undesirable. Lord and Novick (1974) make the following particularly apt comment about reliability:

> . . . maximizing the reliability may sometimes be an undesirable goal. For example, a subset of factual items in an achievement test may yield a more reliable score than the total set of items. This can happen, for example, if the other items involve such hard-to-measure but important traits as reasoning ability and creative thinking. (p. 344)

Secondary Analyses of the Data

An Item Analysis

No treatise on a test is complete without a discussion of an item analysis for that test. The analysis for this test was performed using the Generalized Item and Test Analysis Program (GITAP) (Baker, 1963). This procedure yields the following parameters: item difficulty, biserial correlation, X_{50}, and β.

The item difficulty is the proportion of children responding correctly to an item. The biserial correlation (item-criterion correlation) is obtained by hypothesizing the existence of a continuous latent variable underlying the dichotomy imposed in scoring the item. It is assumed the distribution of measures in the sample for which binary values are given is actually normal but that at some point in the distribution a separation has been made with those cases lying above the point being assigned a score of 1 (correct) and those below a score of 0 (incorrect). One may think of the biserial correlation as a sample measure of association for the item score and the total test score. Under the assumption that the scores are normally distributed and the regression of total test score on the item score is linear, the biserial correlation may be viewed as an estimate of the product moment correlation between the item score and the total test score.

The X_{50} and β are also obtained by hypothesizing the existence of a continuous latent variable underlying the dichotomy imposed in scoring the item. The regression of the hypothetical item score on the criterion is called the

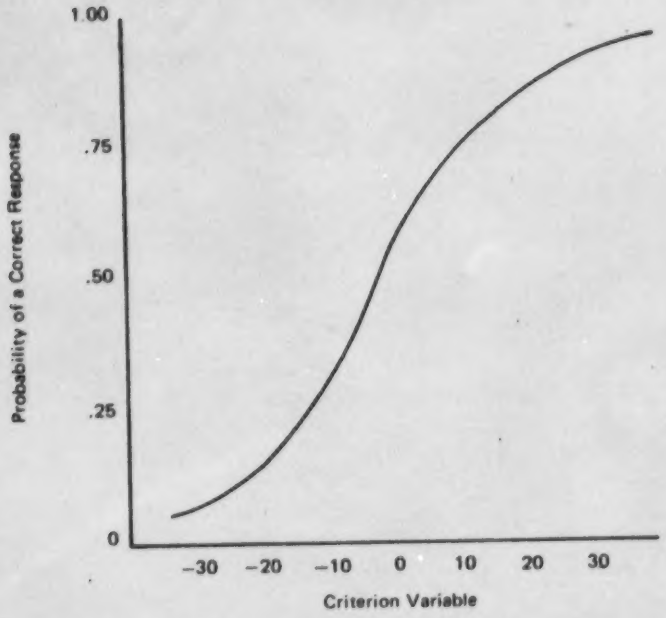


Figure 3. Typical item characteristic curve.

item characteristic curve. The X_{50} is the point on the criterion scale, given in standard deviation units, corresponding to the median of the item characteristic curve. It is the reciprocal of the standard deviation of the item characteristic curve. One may think of β as the slope of the item characteristic curve at the point of X_{50} although this is technically not true. Figure 3 is an illustration of a typical item characteristic curve. The biserial correlation and β are related in the following manner, under the assumption the criterion is normally distributed:

$$\beta = \frac{r_{bis}}{\sqrt{1 + r^2_{bis}}}$$

The generally accepted criteria for a "good item" are β and biserial correlation values of at least .30 (Harris, 1968). Obviously, the higher the value of β, the greater the slope of the item characteristic curve, indicating the item is discriminating more clearly. Thus, an item to which everyone responded correctly would be judged a "poor" item. Although one would like high biserial correlations and a high value for β, it is theoretically possible for them to be too high. If for example, an item had a biserial correlation of 1.00, it would be superfluous to include another item with the same difficulty level

Table 11

Item Parameters

Item	β			r_{bis}		
	C	A	P	C	A	P
1	.04[a]	.35	.45	.04[a]	.33	.41
2	.34	.54	.71	.32	.48	.58
3	.34	.68	.75	.32	.56	.60
4	.28[a]	.44	1.34	.27[a]	.40	.80
5	.32	.50	.66	.30	.45	.55
6	.25[a]	.47	.63	.24[a]	.43	.53
7	.49	.88	1.17	.44	.66	.76
8	.30	.50	1.71	.28[a]	.45	.86
9	.40	.54	1.71	.37	.48	.86
10	.47	.79	.37	.43	.62	.35
11	.29[a]	.51	.65	.28[a]	.45	.54
12	.26[a]	.51	.44	.26[a]	.45	.40
13	.43	-.04[a]	.39	.39	-.04[a]	.36
14	.81	.77	.82	.65	.61	.63
15	.59	.60	.60	.51	.51	.51
16	.63	.60	1.21	.51	.53	.77
17	1.29	.96	.63	.79	.69	.53
18	1.18	.92	.81	.76	.68	.63
19	.98	.38	.20[a]	.70	.35	.19[a]
20	1.12	.56	.21[a]	.75	.49	.21[a]
21	.94	.81	.35	.69	.63	.33
22	.87	.41	.75	.66	.38	.60

[a]Poor questions.

since a subject would respond the same to both items. Similarly, an increase in the value of β, while the values of other parameters remain fixed, could lower the precision of an estimate. This is described as the attenuation paradox (Loevinger, 1954). Therefore, an ideal test should have varying values for β.

Eight of the 66 items on the test had values for β which were less than the desired level of .30. Five of these eight items were comprehension items, one was an application item, and two were problem-solving items. Of these eight items, the value of β for four was between .25 and .29; these values indicate marginal acceptability. All four items were comprehension items. The β value for the comprehension item of superitem 1 (Figure 1) was .04; this is not an unusual value for an item to which 98% of the children responded correctly. The value of β of the application item of superitem 13 was -.04. None of the children who responded correctly to the problem-solving item of this superitem responded correctly to the application item. Therefore, it is not surprising the application item would correlate poorly with the total test score.

The items with low values for β had biserial correlation values of approximately the same size. The additional item with a low biserial correlation was the comprehension item of superitem 8; the biserial correlation re-

Table 12

Means on C, A, and P Scale for Each Cluster

Cluster	Number of subjects	Means		
		C	A	P
1 (−)	91	11.98	6.30	1.59
2 (0)	112	15.96	9.30	2.35
3 (+)	99	17.72	13.17	4.91
4 (++)	15	19.33	17.00	10.01
Total population	317	15.53	10.01	3.30

ported for this item was .28 indicating the item is marginally acceptable. The β and biserial correlation values are given in Table 11.

In conclusion it may be said that all but four of the 66 items had values for β which were acceptable (at least .30) or marginally acceptable (.25 - .29).

The Results of a Cluster Analysis

A clustering procedure was used to discover if there was any structure (natural arrangement of children into homogeneous groups) inherent in the data. Investigators using the test may correlate measures of interest with any of the three scores produced by the test. However, additional insight may be gained by examining the correlations of specific groups of children with similar characteristics identified by the test.

A Wards clustering procedure (Johnson, 1967) was used. The Wards procedure is a maximum method clustering, that is, the value of the clustering is the maximum diameter of the clusters produced. At any step in the clustering process, the distance from an object in cluster j to an object in cluster k is the diameter of the cluster which is the union of clusters j and k. This procedure appeared to produce four clusters.

Cluster 1 consisted of 91 children whose mean comprehension, application, and problem-solving scores were all below the grand means; this group was designated as the (−) group. The mean scores of the 112 children in Cluster 2 were approximately at the mean for all three scores; this group was called the (0) group. Cluster 3 consisted of 99 children whose mean scores were above the grand mean for all three categories of items; this group was called the (+) group. The fourth cluster consisted of 15 children whose application and problem-solving scores were two standard deviations above the mean for the entire group of children; it was impossible for comprehension scores to be two standard deviations above the mean. Cluster 4 was designated the (++) group. The mean scores for each of the clusters are given in Table 12 and the range of the scores within a cluster is shown in Figure 4.

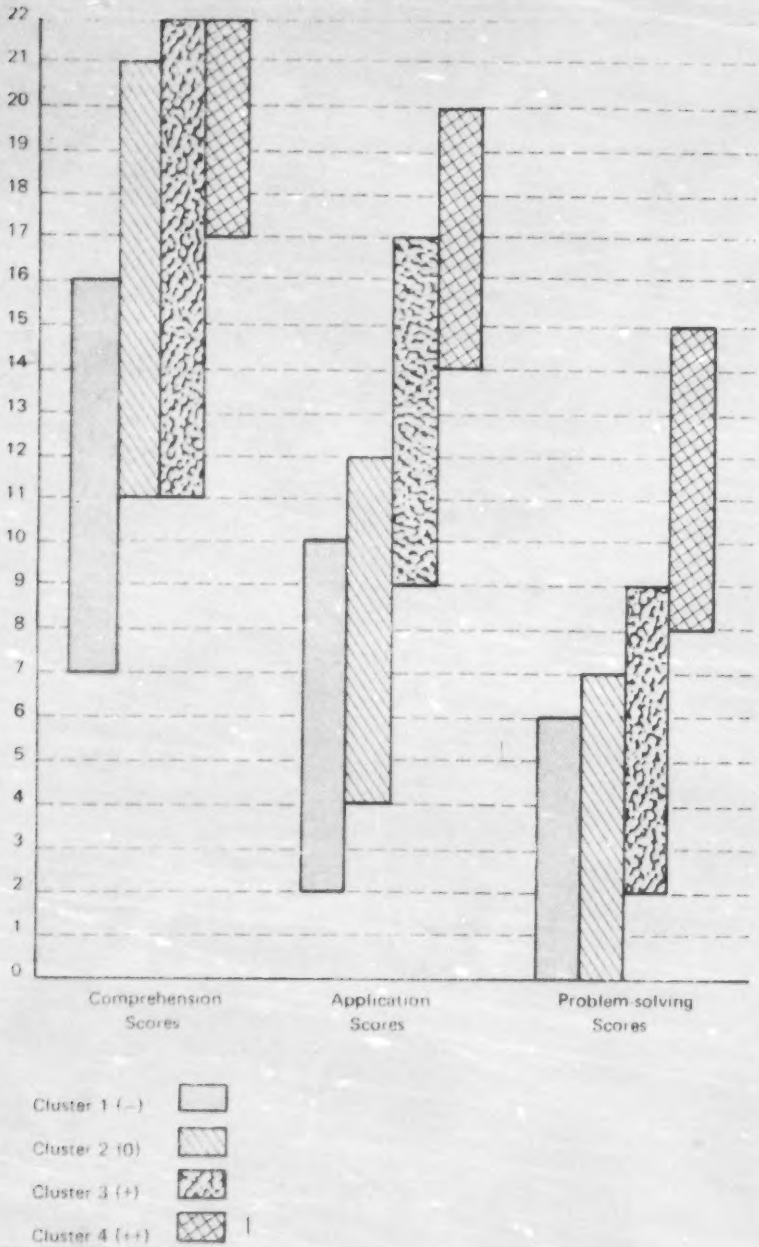


Figure 4. Range of the scores within each cluster.

Conditional probabilities for the items were examined for the children in each of the four clusters. The number of items fitting the model was fewer in the (−) group than in the (++) group; only five superitems were acceptable in the (−) group, indicating that all four conditional probabilities were at least .75. Sixteen superitems were acceptable in the (++) group. Ten items were acceptable or marginally acceptable for Cluster 1, 14 for Cluster 2, 19 for Cluster 3, and 21 for Cluster 4. The conditional probabilities for the superitems in each cluster are contained in Table 13 through Table 16; the number of superitems fitting the categories of Acceptable, Marginally Acceptable, and Unacceptable for each cluster are contained in Table 17.

A superficial explanation for the trend towards more superitems fitting the model as the scores of the children increase may be that the conditional probabilities reflect correct responses to more of the items. That is, if the children are responding correctly to virtually all of the items, then the conditional probabilities would necessarily be close to 1.00. However, the mean problem-solving score for Cluster 4 (++) was 10.01 out of a possible score of 22; this indicates that even the children in the highest category were not responding correctly to the majority of the problem-solving items.

Another explanation for the trend towards more of the items fitting the model as scores increase is the children do not respond as erratically to the multiple choice items in the (+)and (++) groups as in the (−) and (0) groups.

Table 13

Conditional Probabilities for Cluster 1

Conditional probability	P(c | a)	P(c | a)	P(a | p)	P(c ∩ a | p)
	Superitem numbers			
.90 - 1.00	1,2,3,5,8,15, 17,18,22	2,5,10,22	1,2,5,8,13, 17,18,22	5,22
.80 - .89	12,13,14	3	3,9	2
.70 - .79	6,9,20	1,7	11	1,3
.60 - .69	11	9,17	12,14,20	17
.50 - .59	4,10	6,11,12,18	4,6,7	6,9,11,18
.40 - .49	7		19	12
.30 - .39	21	14,20,21	21	20
.20 - .29	19	4		4,7
.10 - .19		8		
.00 - .09	16	13,15,16,19	10,15,16	8,10,13,14, 15,16,19,21

Table 14

Conditional Probabilities for Cluster 2

Conditional probability	P (c | a)	P (c | p)	P (a | p)	P (c ∩ a | p)
	Superitem numbers			
.90 - 1.00	1,2,3,5,8,9, 15,17,20,22	8,17	1,2,5,8,9, 13,14,15,16, 17,18,20	8,17
.80 - .89	6,13,14,18	1,7,9	3,4	1,5
.70 - .79	4,7,12,19,21	2,3,14	6,7,12,21,22	2,3,9,14
.60 - .69	10	4,6,11,18,22	10,11	4,22
.50 - .59	11,16	21		6,7,11,18
.40 - .49			19	
.30 - .39		12		21
.20 - .29		15		12,15
.10 - .19		19		19
.00 - .09		10,13,16,20		10,13,16,20

Table 15

Conditional Probabilities for Cluster 3

Conditional probability	P (c | a)	P (c | p)	P (a | p)	P (c ∩ a | p)
	Superitem numbers			
.90 - 1.00	1,2,5,8,9,11, 14,15,17,18, 20,22	2,5,6,17,20	1,2,3,5,8,9, 10,14,15, 17,18	2,5,17
.80 - .89	3,4,7,12,16,19	1,3,7,9,10,14, 18,20,22	4,12,20,22	1,3,9,10,14,18, 20,22
.70 - .79	6,10,21	11,12,21	6,7,11,21	6,7
.60 - .69		4,8		4,8,11,12
.50 - .59			13,19	21
.40 - .49				
.30 - .39			16	
.20 - .29	13	15		15
.10 - .19		19		
.00 - .09		13,16		13,16,19

Table 16

Conditional Probabilities for Cluster 4

Conditional probability	P (c | a)	P (c | p)	P (a | p)	P (c ∩ a | p)
	Superitem numbers			
.90 - 1.00	1,2,4,5,6,7,8, 9,10,12,13, 14,15,16,17, 18,20,21,22	1,2,3,4,5,6, 7,9,10,11,12, 17,18,21,22	2,4,5,7,8, 9,10,12,13, 14,15,17,18, 19,21,22	2,5,7,9,10, 12,17,18, 21,22
.80 - .89	11	8,14	1,6,11	1,4,6,8,11,14
.70 - .79	19		16	
.60 - .69	3	15,20	20	15,20
.50 - .59		16,19	3	3,16,19
.40 - .49				
.30 - .39				
.20 - .29				
.10 - .19				
.00 - .09		13		13

Table 17

Categorization of the Items on the Basis of Their Conditional Probabilities by Cluster

		Number of items with all four conditional probabilities at that level			
Category	Probability	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Acceptable	.75 - 1.00	5	5	12	16
Marginally acceptable	.50 - .74	5	9	6	5
Unacceptable	.00 - .49	12	8	4	1

Results of a Second Clustering Procedure

A second clustering procedure was performed, this time disregarding the comprehension scores. Two factors prompted the decision to omit the comprehension scores: the conditional probabilities and the length of the test. The mean conditional probability of responding correctly to a comprehension item following a correct application response is .86; if the four lowest conditional probabilities are omitted the mean conditional probability is .90. A child is reasonably certain to respond correctly to a comprehension item after responding correctly to the application item of that superitem.

The second factor influencing the decision to examine the data for structure without the comprehension scores was the length of the test. The test consists of 66 items to be administered within a 45-minute time period. Deleting the comprehension items would release response time for the application and problem-solving items without increasing the total administration time. Deleting the comprehension items is a fairly serious step as these items indicate understanding of the information in the item stem, and incidentally, provide the child with items which are easier at regular intervals. Due to these two factors, it was interesting to see if a clustering procedure would produce meaningful groups when the comprehension items were omitted and if so, to what extent the groups differed from those produced by the initial clustering process.

The initial clustering process produced four groups, three of similar size, 91, 99, and 112, and one small group of 15. The second clustering procedure also appeared to yield four clusters; however, one of these groups was almost one-half of the total group, 147 of the 317 children. Two other groups were approximately the same size, 82 and 72, and there was one small group of 16.

The four new clusters may be designated in a manner similar to those formed in the initial clustering. Cluster A, the largest, consisted of children whose mean application and problem-solving scores were below the grand mean; this group will be designated as the (−*) group. Cluster B consisted of children whose mean application and problem-solving scores were similar to the mean scores for the entire group; this group will be called the (0*) group. Those in Cluster C had mean application and problem-solving scores which were above the mean, this group will be called the (+*) group. Cluster D, again the smallest, consisted of those whose mean application and problem-solving scores were two standard deviations above the mean for the entire group; this group will be designated the (++*) group. The mean scores of the children in the clusters is shown in Table 18 and the range of the scores within a cluster is presented in Figure 5.

The change in the groups from the initial clustering to the second clustering was as follows:

Table 18

Means on A and P Scale for Each Cluster

Cluster	Number of subjects	Means	
		A	P
1 (−°)	147	7.10	1.52
2 (0°)	82	10.71	3.28
3 (+°)	72	13.61	5.44
4 (++°)	16	17.00	10.00
Total population	317	10.01	3.30

1. There were 147 children in Cluster A, 87 of these were in Cluster 1 and 60 in Cluster 2.

2. Cluster B consisted of 82 children; four of these were in Cluster 1, 50 in Cluster 2, and 27 in Cluster 3.

3. All but one of the children in Cluster C were in Cluster 3, the remaining child was in Cluster 2.

4. Fifteen of the 16 children in Cluster D were in Cluster 4, one child was in Cluster 3.

Therefore, it may be said that the comprehension scores had the effect of placing six children in a lower category and 89 in a higher category. This is apparent if one considers the mean scores of the children in the (0) cluster. These scores were slightly below the mean while those in the (0°) cluster are approximately at the mean.

There is one important diffference between the clusters produced using two scores and those using all three; the range of the problem-solving scores within a cluster is smaller when the comprehension scores are omitted than when they are used. That is, the groups are more homogeneous with respect to the problem-solving scores. In the initial clustering procedure, there were children in the (−) group, for example, who had problem-solving scores above the mean. They had responded correctly to fewer application and comprehension items than most, yet at the same time responded correctly to more problem-solving items than most; there were four such children. The range of the problem-solving scores for Clusters 1, 2, 3, and 4 was 7, 8, 8, and 7, respectively. The range for the clusterings in which the comprehension scores were omitted was 3, 6, 7, and 8 for Clusters A, B, C, and D, respectively. This difference may provide slightly different results when other measures are correlated with the scores of children within a cluster.

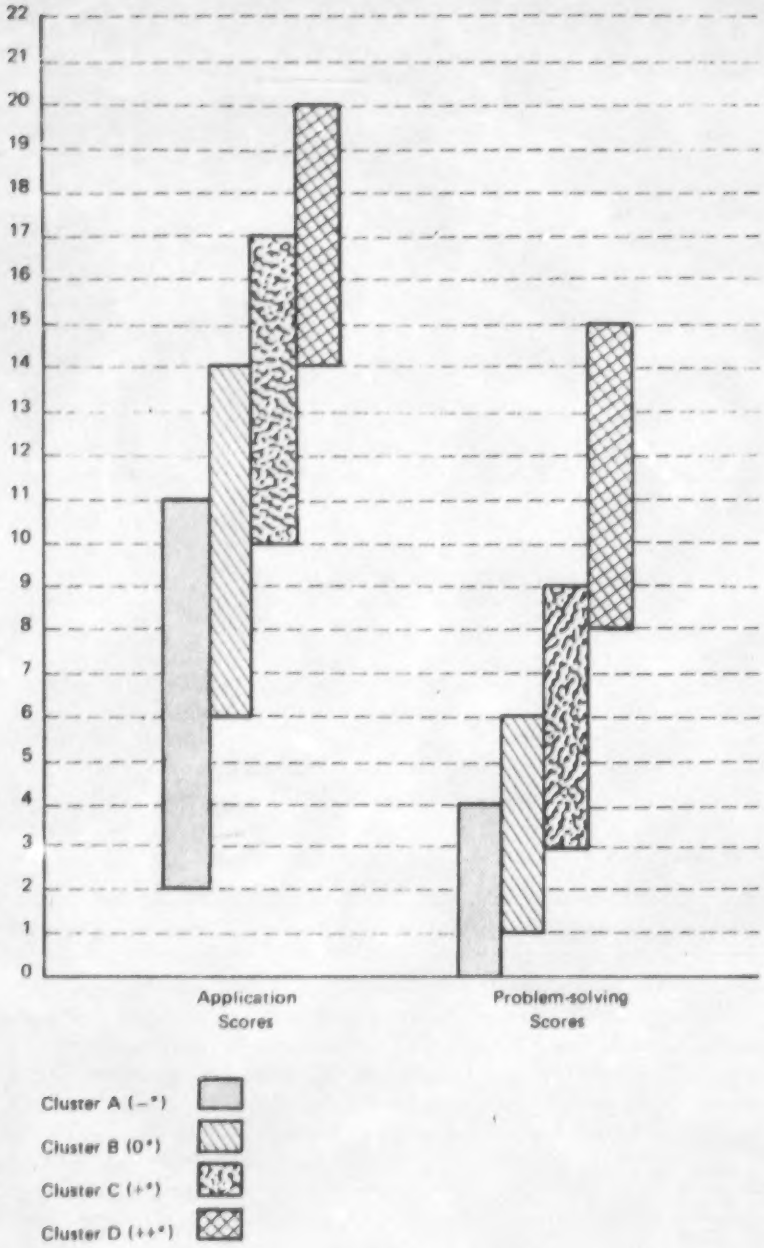


Figure 5. Range of the scores within each cluster.

Table 19

Intercorrelations of the Item

	C	A	P
C	1		
A	.641	1	
P	.431	.667	1

The Complexity of the Items

An insight into the relative complexity of the three types of items is provided by Guttman (1954). Guttman's theory is concerned with studying the order of complexity of a set of variables. The order of complexity is determined by the intercorrelations of the variables. The intercorrelations in a perfect simplex are generated by the law $r_{jk} = a_j / a_k$ where r_{jk} is the correlation between the jth and kth variables and a_j and a_k are the simplex loadings for these variables. If the intercorrelations form a perfect simplex, the variables are said to have a simple order of complexity.

The correlation between the comprehension and application items was .641; it was .667 between the application and problem-solving items, and .431 between the comprehension and problem-solving items. The matrix of intercorrelations is shown in Table 19.

The intercorrelations almost form a perfect simplex; the correlation between the comprehension and the problem-solving scores would have had to be .428 rather than .431 to be a perfect simplex. Consequently, the problem-solving items are more complex than the application items which, in turn, are more complex than the comprehension items. This relationship appears to give further support to the hypothesis that the items on the test fit their definitions.

It has been noted previously that the items are presented in order of difficulty; however, there is a difference between more difficult and more complex. Difficulty in test theory connotes a relationship between group means. Consequently, it is always possible to structure tests to make one type of behavior more difficult than another. Complexity, however, is defined in terms of correlation coefficients and a correlation coefficient is invariant under any linear transformation of scores. Hence, changing group means need not change the correlation coefficients; in fact, the intercorrelations with other tests in a simplex can be essentially the same even if the order of difficulties is reversed. The reason for this is the correlation coefficients depend on the rank order of the people who take the test, and while scores may change, the rank order of the subjects does not change. Therefore, Pearsonian coefficients usually do not vary much since they are closely related to rank correlations.

Thus, one may state that the items which comprise the superitems are not only increasing in difficulty, they are also increasing in complexity.

Summary

The purpose of the study was to develop a test of mathematical problem-solving behavior which provided information about the child's mastery of the prerequisites of each problem-solving question. To provide this additional information, each problem-solving question was preceded by two other questions, all related to the same item stem. One question assessed the child's understanding of the information contained in the item stem and a second assessed knowledge of an underlying concept or process of the problem-solving question. The first was referred to as the comprehension question and the second, the application question.

To determine whether asking multiple questions on the same unit of information affected response to the items, the three item types were administered alone, in pairs, and all three together. Means for each of the three scales, Comprehension, Application, and Problem-solving, were compared across the four instruments containing each scale. A post hoc procedure identified two significant differences among all 18 differences considered.

Two hypotheses were advanced to account for the significant differences: (a) the children did not have the same amount of time to respond to a particular group of items on all of the tests containing that group of items, and (b) asking multiple questions on the same unit of information affects the response to the question. The conclusion was that the significant differences were probably the result of administration times; that is, the children did not have the same amount of time to respond to a particular category of items on each test containing those items and this affected their scores.

Conditional probabilities were computed for each superitem to ascertain if the comprehension item was assessing a real prerequisite of the application item and if both comprehension and application items were assessing actual prerequisites of the problem-solving item. A superitem was deemed acceptable if all four of the conditional probabilities were at least .75; 10 of the 22 superitems were in this category with an additional six marginally acceptable. If one of the four conditional probabilities was less than .50, the item was considered unacceptable; five superitems were in this category. Two of these five superitems were believed to be in this category for reasons other than failure to fit the test model.

The only indicator of validity that could be obtained was the content validity of the test as measured by a panel of judges interested in problem-solving research. The judges' independent classification of the items was compared to the actual classification of the items on the test and a measure of association was computed between these two classifications; the computed measure was .78.

Two reliability estimates were computed for the test, a Hoyt reliability estimate under the assumption the items were independent and a generalized

KR-20 reliability estimate in the event the items were not independent. The Hoyt estimate was .84 and the KR-20 was .79. Due to the conclusion the items were independent, the .79 estimate was considered conservative and the Hoyt estimate of .84 was believed to better represent the reliability of the test.

A cluster analysis was used to discover if there was any structure inherent in the data. A first cluster analysis appeared to produce four groups of children. One group had mean scores which were below the mean, one at the mean, one above the mean, and one considerably above the mean scores for the entire group. The results of a second clustering, produced without the comprehension scores, were similar to those formed by considering all three scores with the exception that the range of the problem-solving scores within a cluster was smaller when the comprehension scores were omitted.

Conditional probabilities for the superitems were examined for each of the four clusters formed by considering all three scales. The number of superitems fitting the model was fewer in the (−) cluster than in the (++) cluster. This may have reflected the children in the (++) group responding more consistently to the multiple choice items than those in the (−) group.

A Guttman analysis indicated that the three categories of items were in order of complexity, that is, the problem-solving items were more complex than the application items which were more complex than the comprehension items. Thus, the items comprising the superitems are not merely in order of difficulty, but also assess increasingly complex behavior.

Conclusion

The results of the study support the contention that a test composed of superitems is a viable form for assessing problem-solving behavior. The superitem test produces three scores with which to correlate other measures and scores of groups of children formed by a clustering procedure.

Chapter 9

Mathematical Problem-solving Performance and Intellectual Abilities of Fourth-grade Children

Ruth Ann Meyer

The purpose of this study was to investigate relationships between mathematical problem-solving performance and intellectual abilities. More specifically, the investigator attempted to identify a structure of mathematical problem-solving performance.

Background

This study's inception and design are attributed primarily to *A Structure of Concept Attainment Abilities Project* (CAA) (Harris & Harris, 1973). The CAA study was conducted at the Wisconsin Research and Development Center for Cognitive Learning during 1970 and 1971 to determine a structure of concept attainment abilities. Batteries of cognitive abilities reference tests and tests to measure attainment of mathematics, social studies, science, and language arts concepts were administered by the CAA staff to samples of fifth-grade males and females. Through factor analysis, a basic cognitive abilities structure and relationships between concept learning and cognitive abilities in the four selected school subjects were identified. Harris and Harris (1973) summarized the results:

> We conclude that seven latent cognitive abilities underlie the test batteries that were studied and that these are the same for both boys and girls. The seven abilities are: Verbal, Induction, Numerical, Word Fluency, Memory, Perceptual Speed, and Simple Visualization. The first six are the seven Primary Mental Abilities of the Thurstones. The seventh is similar to the Thurstones' Closure One but we prefer to call it Simple Visualization. (p. 169)

Furthermore, the CAA staff found that:

1. Achievement in science and social studies was related to three abilities — Verbal, Induction, and Memory.
2. Achievement in language arts and mathematics was related to three additional abilities — Numerical, Word Fluency, and Memory.
3. Two abilities—Perceptual Speed and Simple Visualization—seemed not to be related to achievement in these four subject-matter fields. (p. 195)

Relevant Literature

There are many studies such as Balow (1964), Beldin (1960), Johnson (1949), Linville (1969), Norman (1950), Thompson (1967), and Treacy (1944), which demonstrate the influence of a single element such as reading comprehension, vocabulary, or computational ability upon success in mathematical problem solving. However, few investigations show relationships which may exist between a combination of elements or abilities and mathematical problem solving. Consideration of a structure of intellectual abilities related to successful mathematical problem solving has been rare.

Studies of structures of intellectual abilities related to problem solving have primarily investigated mathematical ability. These studies provide some insights into mathematical problem solving as their batteries of tests generally include a problem solving or application test. An example was the investigation by Very (1967), who administered a battery of 30 tests to 335 university students. All tests were chosen to measure abilities considered pertinent to mathematical ability. Data for the total group, for males only, and for females only, were subjected to factor analysis by principal component procedures. For all three groups, Verbal, Numerical, Perceptual Speed, Spatial Ability, and General Reasoning factors were found. The General Reasoning factor, Arithmetic, Deductive, and Inductive Reasoning factors were isolated for males only. Although three additional reasoning factors emerged for females, Very found the factors difficult to define.

The principal aim of the investigations by Werdelin (1958) was to analyze the structure of the problem-solving aspect of mathematical ability. Numerical, Verbal, Visual, Deductive, and General Mathematical Reasoning factors were found in both his Alpha and Beta studies. After reanalysis of the data in 1966, Werdelin commented:

> Problem solving in mathematics depends primarily on the general reasoning factor R, according to the results of this study. Only to a somewhat smaller extent are factors like the deductive reasoning factor D and the numerical factor N of importance. This is a result which is closely related to the very nature of mathematics problem solving.
>
> A problem is a task which involves several elements which shall be combined in the solution. The elements may be taken from various fields, such as the verbal one, the numerical one, the visual-perceptual one, etc. Therefore, it is to be expected that these problems are loaded on the R factor as it is interpreted in the above.
>
> Our having rotated the two studies to a common structure has enabled us to conform the existence of the five factors. Furthermore, it has aided us in interpreting these. There are several questions which need to be further studied, however. The nature of factors like D and R is still little known and their fields of definition are largely unknown. The number of

factors in the visual-perception field and the reasoning should be studied, and so on. The main result of the present study is probably our having founded a platform on which to build a larger structure. (p. 13)

Other investigations of problem-solving structures were conducted by graduate students at the Catholic University of America (Campbell, 1956; Donohue, 1957; Edwards, 1957; Emm, 1959; Engelhard, 1955; Kliebhan, 1955; McTaggart, 1974). Tests of problem solving and other tests believed to be related to problem solving, were administered to groups of fifth-, sixth-, and seventh-grade males and females. Verbal and Arithmetic factors were identified for each of the six groups. In addition, Campbell (1956) found, for sixth-grade males, a factor which involved comparison of data prior to problem solving. Donohue (1957) found an Approach-to-Problem-Solving factor for seventh-grade males and females, Emm (1959) identified a Spatial factor for fifth-grade males, and McTaggart (1974) found another Verbal factor for fifth-grade females.

These factor analytic studies of mathematical ability and problem solving, as well as the CAA Project, suggested the existence of a stable intellectual structure of Verbal, Numerical, Reasoning, Spatial, Perceptual Speed, and Memory factors. How each of these factors related to mathematics achievement was not clear; but, significantly, one of the reasoning factors of each of the mathematical ability and problem-solving studies had been determined primarily by mathematics tests. It was the purpose of the present factor analytic study to investigate the relationship between these stable intellectual factors and mathematical ability, particularly mathematical problem solving.

Procedures

Subjects

The subjects of this study were 179 fourth-grade children from Wisconsin, Illinois, and New York. Participation was determined by: (a) enrollment in *Developing Mathematical Processes* (DMP), a K-6 elementary mathematics program developed by the Analysis of Mathematics Instruction Project of the Wisconsin Research and Development Center for Cognitive Learning (Romberg, 1976; Romberg, Harvey, Moser, & Montgomery, 1974, 1975, 1976); (b) fourth-grade level; (c) geographic area; and (d) willingness of principals and teachers to have their students included in the study.

To ensure similarity in experiential background for the sample, the investigation was restricted to fourth-grade children who were studying DMP. Since, at that time, only a few pilot schools were using DMP materials beyond fourth grade, it would have been difficult to procure a sample of 200 children at any higher grade. The mathematical problem-solving test was designed for children in at least fourth grade. The geographic area constraint was primarily for the convenience of the investigator.

Instruments

Twenty tests were administered to the sample in this study. Of these tests, 19 were "reference" tests for intellectual abilities and the remaining test was a mathematical problem-solving test constructed by Romberg and Wearne (Wearne, 1976). The Romberg-Wearne test was designed to yield three scores: a comprehension score, an application score, and a problem-solving score. To accomplish this, the test was composed of groups of items called *superitems;* each of these superitems contained an item stem, a comprehension item, an application item, and a problem-solving item. This Romberg-Wearne problem-solving instrument and its superitems were described in Chapter 8. An example of a superitem is given here:

A parking lot has room for 8 row of cars with 9 cars parked in each of those rows. (Item stem)

The parking lot has room for the same number of cars in each of 8 rows. (Comprehension item)

How many cars can be parked in the parking lot? (Application item)

In another parking lot, trucks are parked. Each truck takes the space of 3 cars. There are 12 trucks in the parking lot and it is completely full. If there were 4 rows in the parking lot, how many cars could be parked in each row? (Problem-solving item)

Although the primary objective of this study was to examine performances of children in problem situations similar to those in the problem-solving questions of the Romberg-Wearne test, the test also provided information about prerequisite computation skills and mathematics concepts for the problem-solving questions. Therefore, the three measures of the Romberg-Wearne test, a Comprehension score, Application score, and Problem-solving score, were used in all analyses of this study.

Table 1 lists the other 19 tests administered to the sample. This table indicates the intellectual abilities hypothesized for the respective reference tests and gives the source of each test. Intellectual structures identified in past factor analytic studies of mathematical ability or mathematical problem solv-

Table 1

Intellectual Abilities Hypothesized for the Population Sample, the Respective Reference Tests, and their Sources

Intellectual abilities	Reference tests	Sources
Verbal	Picture Group Name Selection (12)[a]	Constructed by CAA[b] staff
	Word Group Naming (19)	Constructed by CAA staff
	Remote Class Completion (14)	Adapted from Waddle Test
	Vocabulary (18)	Iowa Tests of Basic Skills
Induction	Letter Classification (4)	Constructed by CAA staff
	Number Classification (6)	Constructed by CAA staff
	Figure Matrix (1)	Sheridan Psychological Services
	Number Exclusion (7)	Constructed by CAA staff
Numerical	Mathematics Computation (5)	Constructed by Romberg
	Number Series (8)	Constructed by CAA staff
	Seeing Trends (15)	Constructed by CAA staff
Word fluency	Omelet (9)	Constructed by CAA staff
	Spelling (17)	Iowa Tests of Basic Skills
Memory	Remembering Classes: Members (13)	Constructed by CAA staff
	Picture Class Memory (11)	Constructed by CAA staff
Perceptual speed	Identical Pictures (3)	ETS Kit of Reference Tests
	Perceptual Speed (10)	PMA 4-6 Test Battery (SRA)
Simple visualization	Gestalt Completion (2)	Constructed by CAA staff
	Spatial Relations (16)	PMA 4-6 Test Battery (SRA)

[a]Numbers in parentheses represent the alphabetical order of the tests. This order is used when describing the tests.

[b]CAA refers to *A Structure of Concept Attainment Abilities Project* (Harris & Harris, 1973).

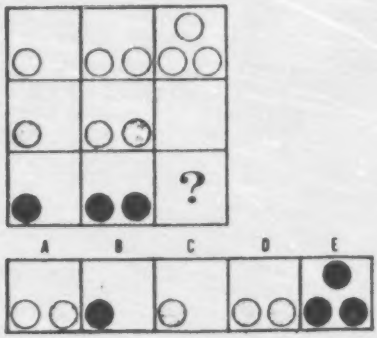
ing, and the CAA Project (Harris & Harris, 1973) suggested the hypothesized structure for the present study.

All but one of the reference tests were selected from those used in the CAA study. Eleven of those tests, Gestalt Completion, Letter Classification, Number Classification, Number Exclusion, Number Series, Omelet, Picture Class Memory, Picture Group Name Selection, Seeing Trends, Word Group Naming, and Remembering Classes: Members, were developed by the staff of the CAA Project (Harris & Harris, 1973). Two of the tests, Perceptual Speed and Spatial Relations, are a part of the Primary Mental Abilities 4-6 Test Battery (Science Research Associates, 1962). Two other tests were from the Iowa Tests of Basic Skills (Lindquist & Hieronymus, 1964); they were Spelling and Vocabulary. Identical Pictures is a part of the ETS Kit of Reference Tests (French, Ekstrom, & Price, 1969a, 1969b), and Figure Matrix was developed by Sheridan Psychological Services, Inc. (1969). The Remote-Class Completion test (Harris & Harris, 1973) was adapted from the Wad-

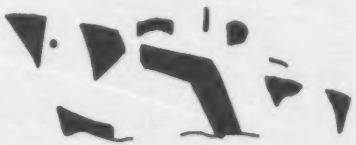
dle Test. The one test used which was not in the CAA battery was Mathematics Computation (Romberg, 1975). The investigator attempted to select from the CAA battery those tests which she hypothesized were related to problem solving. Also, since this was to be a factor analytic study, at least two reference tests were included for each hypothesized ability. A brief description of each of the tests chosen is given in the next section.

Description of Reference Tests for Cognitive Abilities

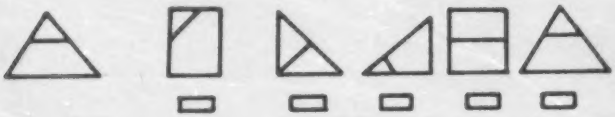
1. *Figure Matrix.* In this test the subject is to infer two spatial relations (across and down) and combine them. Then the figure that belongs in the cell with the question mark is selected from five choices. Example:



2. *Gestalt Completion.* This test involves naming an object from a partially obliterated picture. Example:



3. *Identical Picture.* In this test the subject selects from five choices a figure identical to a given one. Example:



4. *Letter Classification.* In each item of this test the subject is to infer a class from three given examples. Then a fourth example of the class is selected from three choices. Example:

A B C C	1. B A B C
C D A A	2. A D B B
B D A A	3. A A C B

5. *Mathematics Computation* (Romberg, 1975). This test consists of the following types of problems: addition, subtraction, place value, ordering, finding the missing number, and representing parts of a whole.

6. *Number Classification.* In this test, similar to Letter Classification, the subject is to examine the structure and form of three examples and infer a class to which all three belong. Then another example of that class is selected from five choices.

7. *Number Exclusion.* This test parallels Number Classification, but the task required is exclusion rather than classification. Given four examples, the subject is to infer a class that includes three of them, and to indicate the example that is excluded from that class. Example:

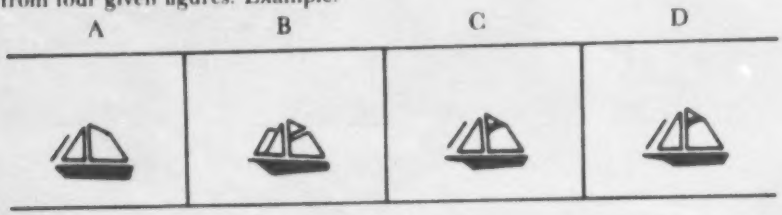
A. 5 B. 75 C. 750 D. 885

8. *Number Series.* Numbers forming a series are given in this test. The subject must infer a quantitative rule and indicate which of five choices would come next in the series. Example:

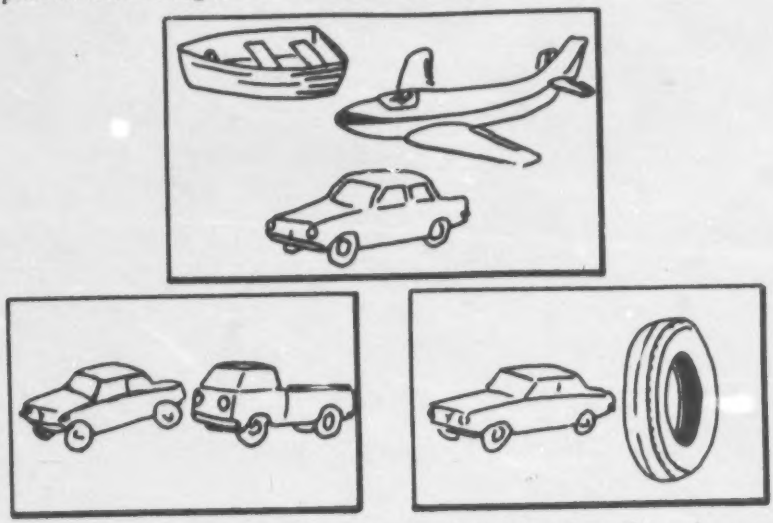
2 8 14 20 A. 16
B. 20
C. 22
D. 24
E. 26

9. *Omelet.* In this test words are given with the letters in scrambled order. The subject is to identify each word and spell the word correctly.

10. *Perceptual Speed.* This test requires circling two identical pictures from four given figures. Example:



11. *Picture Class Memory.* In this test the subject studies 10 sets of 3 pictures. The three pictures in each set are examples of a class. The subject infers the class, remembers it, and then judges whether or not 20 sets of 2 pictures each belong to a class that was studied. Example:



12. *Picture Group Name Selection.* In this test three pictured examples of a class are given. The subject is to infer the class and select the best name for the class. Example:

All are:
A. animals
B. brown animals
C. dogs

13. *Remembering Classes: Members.* For this test the subject studies 10 sets of 3 words. Immediately following the study period, the subject indi-

cates whether or not each of 20 sets of 2 words belongs to a class that was studied. Example:

A.	daisy	I.	daisy
	rose		pansy _____
	poppy	II.	daisy
			grass _____

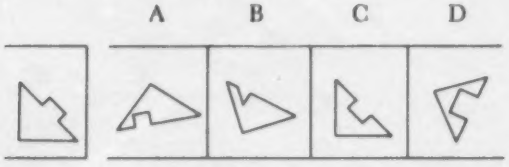
14. *Remote Class Completion.* In this test the subject is to produce a fourth word that goes with three given words. The given words all go together in some way, but the class is a remote one. Example:

America eye hawk _____

15. *Seeing Trends.* In each item of this test four examples are given. The subject infers a rule based on number of letters, alphabetic position of letters, etc. Using this rule, the subject places the word, given in parentheses, in its proper serial position. Example:

all ____ boy ____ cage ____ (dot)
A B C

16. *Spatial Relations.* From four choices the subject chooses the figure that would complete a given figure to form a square. Example:



17. *Spelling.* In each item of this test the subject is to select the misspelled word if there is one, or select "no mistakes" if each of four words is spelled correctly.

18. *Vocabulary.* In each item the subject is to select from four choices a synonym for the underlined word in a phrase.

19. *Word Group Naming.* In each item of this test four examples of a class are given. The subject must supply a name for the class. Example:

tepee	beehive	
		All are _____
nest	igloo	

Methodology

One method for studying relationships between variables is intercorrelation analysis. However, a large number of variables makes the task of explaining all of the resulting intercorrelations nearly hopeless. Factor analysis provides techniques for summarizing relationships between variables, thereby

making interpretations easier. Butcher (1968) commented about factor analysis:

> This is a powerful mathematical technique for unravelling a complex pattern of overlapping influences, and is in many ways ideally suited to provide an answer to the questions that have been asked about the structure of human abilities. Indeed, the views of psychologists at the present time have been more strongly influenced by the results of factor-analyzing test scores than by any other approach. (pp. 42-43)

Since the primary aim of this study was to determine a structure of mathematical problem-solving performance by investigating relationships among a large number of variables, factor analysis was deemed appropriate.

Because factor analysts adhere to different theoretical bases, many factor analytic procedures have emerged. Moreover, when these different methods are used with a given set of data, different factor structures may result. Because of this indeterminacy, the conservative approach to factor analysis of Harris and Harris (1973) was used in this study.

Harris and Harris (1973) used three initial factor methods: Alpha (Kaiser & Caffrey, 1965); Harris R - S^2 (Harris, 1962); and Unrestricted Maximum Likelihood Factor Analysis (UMLFA) (Jöreskog, 1967). Kaiser's normal varimax procedures (Kaiser, 1958) were used in the present study to obtain orthogonal solutions for each of the three initial solutions. For each of the sets of orthogonal common factors, two derived oblique solutions, Independent Cluster and A′A Proportional to L, were derived by procedures used by Harris and Kaiser (1964). It was necessary to procure both oblique solutions for the data as it was impossible to predict which results would be more interpretable.

Results

Means, Standard Deviations, and Reliabilities

The Generalized Item and Test Analysis Program (GITAP) (Baker, 1969) was used to obtain means, standard deviations, and Hoyt analysis of variance reliability estimates for each of the 19 reference tests for intellectual abilities and the three parts of the Romberg-Wearne Mathematical Problem-solving Test. These statistics are presented in Table 2.

The Hoyt reliability estimates for the reference tests were generally good. Ten of the estimates were equal to or greater than .80; two estimates were greater than .90. The reliability of the Application part of the Romberg-Wearne test was .69; however, the reliabilities of the Comprehension and Problem-solving parts were relatively low, .48 and .59, respectively.

Table 2

Means, Standard Deviations, and Reliability Estimates for Test Scores

	Test	Items	Mean	Standard deviation	Hoyt reliability
1	Figure Matrix	20	8.92	3.91	.74
2	Gestalt Completion	20	12.23	3.66	.75
3	Indentical Picture	48	26.61	9.19	.95
4	Letter Classification	20	13.78	3.37	.72
5	Mathematics Computation	54	40.51	8.12	.89
6	Number Classification	30	24.07	5.90	.91
7	Number Exclusion	20	13.88	4.00	.81
8	Number Series	20	12.84	4.15	.81
9	Omelet	20	10.32	5.01	.88
10	Perceptual Speed	40	27.44	6.58	.89
11	Picture Class Memory	20	15.45	3.20	.78
12	Picture Group Name Selection	20	12.08	3.07	.63
13	Remembering Classes: Members	20	13.85	3.43	.71
14	Remote Class Completion	25	12.75	4.08	.76
15	Seeing Trends	20	11.85	3.79	.73
16	Spatial Relations	25	15.94	4.18	.76
17	Spelling	38	14.11	6.95	.87
18	Vocabulary	38	24.40	7.28	.89
19	Word Group Naming	20	12.18	4.27	.80
20	Comprehension	19	13.52	2.41	.48
21	Application	19	9.72	3.33	.69
22	Problem Solving	19	3.47	2.40	.59

Note. Number of subjects is 179.

Single-battery Factor Analyses

To examine the relationships between mathematical problem-solving performance and intellectual abilities, the 19 intellectual ability measures and three problem-solving scores were combined into one matrix for single-battery factor analyses. The intercorrelations of these 22 variables (Matrix B) are given in Table 3.

After finding orthogonal and oblique rotations of the Alpha, Harris R-S^2, and Unrestricted Maximum Likelihood initial factor solutions of Matrix B, an interpretation strategy of Harris and Harris (1971) was applied to the three orthogonal and three A′A Proportional to L oblique solutions. The A′A Proportional to L oblique solution was more easily interpreted for this particular data than was the Independent Cluster solution.

This interpretation strategy attempts to determine factors that are robust with respect to method—factors which tend to include the same variables across methods. A variable was considered relevant to a factor if it had a coefficient greater than .30 (absolute) on that factor. A comparable common factor was defined as having two or more of the same relevant variables on at least four of the six derived solutions.

Table 3

Intercorrelations of Variables: Matrix B

Tests	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1 Figure Matrix																					
2 Gestalt Completion	10																				
3 Identical Picture	17	32																			
4 Letter Classification	46	01	10																		
5 Mathematics Completion	36	16	27	32																	
6 Number Classification	32	14	22	29	38																
7 Number Exclusion	34	08	16	45	49	43															
8 Number Series	49	19	16	41	49	39	43														
9 Omelet	33	34	30	30	52	40	34	41													
10 Perceptual Speed	15	17	54	10	32	06	26	20	11												
11 Picture Class Memory	39	09	20	25	28	32	28	28	31	25											
12 Picture Group Name Selection	48	20	07	33	39	21	27	19	30	19	39										
13 Remembering Classes: Members	37	01	08	31	34	29	29	35	29	26	38	41									
14 Remote Class Completion	39	17	04	32	46	30	22	44	48	18	29	44	31								
15 Seeing Trends	22	17	24	19	29	14	20	42	30	24	00	05	08	20							
16 Spatial Relations	49	17	19	33	35	26	35	48	32	22	40	42	37	30	18						
17 Spelling	31	21	12	29	52	35	25	47	59	08	17	35	39	54	26	23					
18 Vocabulary	39	20	10	41	49	39	29	46	52	12	32	56	47	57	22	29	63				
19 Word Group Naming	52	23	17	47	49	40	35	54	53	21	38	60	45	60	29	45	52	71			
20 Comprehension	36	-01	14	34	54	25	29	41	34	20	23	40	38	40	23	29	42	43	48		
21 Application	45	12	15	46	57	38	38	56	54	22	33	54	45	46	32	38	53	60	66	68	
22 Problem Solving	46	08	14	35	47	26	35	45	47	22	25	34	32	41	35	29	37	36	43	48	60

Note. Decimal points have been omitted.

The Harris and Harris interpretation strategy yielded six comparable common factors. Table 4 gives the loadings of the variables relevant to the respective comparable common factors. Those variables which had loadings greater than .30 on at least four of the derived solutions are given in capital letters.

Comparable Common Factor 1 (B-CCF 1) appeared to be a Verbal factor combining Word Fluency and Verbal Comprehension. Comparable Common Factor 2 (B-CCF 2) was classified as Induction of classes employing symbolic content, and Comparable Common Factor 3 (B-CCF 3) appeared to be a Numerical factor. Comparable Common Factor 4 (B-CCF 4) was readily identified as a Perceptual Speed factor, and Comparable Common Factor 5 (B-CCF 5) was an Induction factor employing verbal semantic, pictorial semantic, or figural content. Last, Comparable Common Factor 6 (B-CCF 6) appeared to be a factor specific to mathematics. This factor was determined primarily by the three scores of the Romberg-Wearne Problem-solving Test.

In addition, Application, Part II of the Romberg-Wearne Problem-solving Test, had small loadings on the orthogonal Harris R-S^2, orthogonal UMLFA, and Alpha oblique solutions of B-CCF 1 (a Verbal factor). Comprehension, Part I of the Problem-solving Test, had a small loading for the two derived UMLFA solutions of B-CCF 3 (a Numerical factor). Furthermore, Application had small loadings for all three orthogonal solutions of B-CCF 5 (Induction), and Comprehension had one small loading for the Alpha orthogonal solution of B-CCF 5.

Frequency Responses for the Problem-solving Test

Table 5 shows that comprehension of the information given in the item stem and mastery of the prerequisite mathematics concept or skill did not guarantee success with the Problem-solving question. For instance, 131 subjects appeared to comprehend data given in superitem 10 and 105 multiplied 8×9 correctly, but only 19 found the correct answer to the Problem-solving question.

Generally, the highest number of correct responses was on the Comprehension questions, the next highest on the Application questions, and the lowest on the Problem-solving questions. Three exceptions were the Application and Problem-solving scores for superitems 4, 7, and 17. In superitem 4, children confused the concepts of perimeter and area in the Application question. In superitem 7, a few children seemed to know the meaning of average, but they could not compute the average of three numbers. The expression $2(n) + 100$ caused children to have difficulty with the Application question of superitem 17.

Discussion and Conclusions

The six comparable common factors (Verbal, two Induction, Numerical, Perceptual Speed, and General Mathematics) resembled the factors of the

Table 4

Comparable Common Factors for Matrix B

	Test	Orthogonal[a]			Oblique[a]		
		A	H	U	A	H	U
	Comparable Common Factor 1 (B-CCF 1)						
9	Omelet	63	47	61	62		51
14	Remote Completion	42	63	54	53	54	38
17	Spelling	60	67	71	69	54	63
18	Vocabulary	53	71	72	67	59	61
19	Word Group Naming	44	58	55	53	41	34
2	Gestalt	45		31	42		
5	Mathematics Computation		41	41			37
6	Number Classification			35			35
7	Number Exclusion						37
8	Number Series		31	32			
12	Picture Group Name Selection		40	32	37		
13	Remembering Classes: Members		36				
20	Comprehension		34				
21	Application		44	42	32		
	Comparable Common Factor 2 (B-CCF 2)						
4	Letter Classification	43	34		38	39	
6	Number Classification	47	51	31	50	50	
7	Number Exclusion	63	60	94	65	63	88
1	Figure Matrix	38					
5	Mathematics Computation	31	39	31			
8	Number Series	42			38		
16	Spatial Relations	35					

Table 4 (Continued)

	Test	Orthogonal[a]			Oblique[a]		
		A	H	U	A	H	U
	Comparable Common Factor 3 (B-CCF 3)						
8	Number Series		41	42		34	50
15	Seeing Trends		56	50		54	49
1	Figure Matrix						45
9	Omelet			36			35
22	Problem Solving			46			42
16	Spatial Relations						42
	Comparable Common Factor 4 (B-CCF 4)						
3	Identical Pictures	72	64	99	72	65	99
10	Perceptual Speed	68	69	53	68	69	50
2	Gestalt	33			32		
	Comparable Common Factor 5 (B-CCF 5)						
1	Figure Matrix	47	62	64	39	51	41
11	Picture Class Memory	57	51	52	54	38	48
12	Picture Group Name Selection	68	56	64	55	54	63
13	Remembering Classes: Members	51	42	45	43		42
16	Spatial Relations	46	60	60	39	49	39
19	Word Group Naming	53	46	57		34	46
4	Letter Classification		37	42			
8	Number Series		42	45			
20	Comprehension	32					
21	Application	40	33	41			
14	Remote Completion	38		36			

Table 4 (Continued)

	Test	Orthogonal[a] A	Orthogonal[a] H	Orthogonal[a] U	Oblique[a] A	Oblique[a] H	Oblique[a] U
18	Vocabulary	47		41			
	Comparable Common Factor 6 (B-CCF 6)						
5	Mathematics Computation	49	39	37	33	48	33
20	Comprehension	65	60	70	59	68	72
21	Application	67	59	56	51	61	53
22	Problem Solving	58	48	38	49	46	41
4	Letter Classifications	32					
8	Number Series	44					
14	Remote Completion	40					
15	Seeing Trends	41					
17	Spelling	45					
18	Vocabulary	38					
19	Word Group Naming	40					
	Factors Specific to Single Initial Solutions						
2	Gestalt		52			45	
9	Omelet		40			40	

Note. Includes coefficients greater than .30 (absolute). Decimals have been omitted.

[a]A = Alpha; H = Harris R-S^2; U = UMLFA.

Table 5

Number of Correct Responses on Each Part of the 19 Superitems

Part	Superitem																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Comprehension	162	158	171	128	168	95	34	86	163	131	132	125	141	162	101	144	101	104	139
Application	133	134	151	38	115	37	8	78	125	105	105	102	116	63	50	99	47	102	130
Problem-Solving	81	12	62	52	32	19	13	35	70	19	37	30	23	6	6	11	50	26	43

Note. There were 179 subjects.

hypothesized structure for this study, although there were differences. The hypothesized factors Word Fluency, Simple Visualization, and Memory were not isolated. The two reference tests for Word Fluency, which were Omelet and Spelling, helped to determine the Verbal factor for the sample. Gestalt Completion, one of the reference tests for Simple Visualization, had small loadings on the Alpha solutions of B-CCF 1 (Verbal) and B-CCF 4 (Perceptual Speed). Spatial Relations, the other reference test for Simple Visualization, and the two reference tests for a Memory factor, Picture Class Memory and Remembering Classes: Members, had significant loadings on all but one of the derived solutions of B-CCF 5. Therefore, induction seemed to be more important than remembering for the two Memory tests and more important than visualizing for the Spatial Relations test.

Comprehension, Application, Problem Solving, and Mathematics Computation determined a General Mathematics factor. This factor resembled Very's Arithmetic Reasoning factor (1967), Werdelin's General Mathematical Reasoning factor (1958, 1966), and the Arithmetic factor identified by the series of studies conducted by graduate students at the Catholic University of America from 1956-1959. In all of these studies, including the present one, a factor specific to mathematics emerged. However, while in past studies this factor was determined by mathematical reasoning, it appeared to be determined by mathematics concepts in the present study.

The loadings of Problem Solving, Part III of the Romberg-Wearne test, on both the orthogonal and oblique UMLFA solutions of B-CCF 3, suggested some relationship between Problem Solving and Numerical Ability. The three small loadings of Application on B-CCF 1 suggested that applications are related slightly to Verbal ability. The other small loadings of Comprehension and Application are probably of little consequence.

Further Analyses

All analyses for this study were for males and females combined. Since the investigator was also interested in any sex-related differences in mathematical problem-solving performance and intellectual structure, the data were reanalyzed for males and females separately.

The *t*-tests employed demonstrated significant sex-related differences for only two of the intellectual variables, Spatial Relations ($p < .01$) and Picture Group Name Selection ($p < .03$). Both were in favor of males. There were no significant sex-related differences for the three scores of the Romberg-Wearne Mathematical Problem-solving Test. However, factor analytic procedures resulted in different structures for males and females.

Five comparable common factors were identified for males and six comparable common factors emerged for females. The factors for males were:

(a) Verbal Ability and Word Fluency, (b) Induction of classes employing symbolic or figural content, (c) Perceptual Speed, (d) Problem Solving, and (e) Mathematics Concepts. The factors for females included: (a) Verbal Ability, (b) Induction of classes employing pictorial, figural, or verbal content, (c) Numerical Ability, (d) Perceptual Speed, (e) a Fluency factor employing either words or numbers, and (f) General Mathematics.

Significantly, the three measures from the problem-solving test resulted in two mathematics factors emerging for males and only one for females. The Problem-solving factor for males was determined primarily by the Problem-solving questions of the Romberg-Wearne test together with the reference tests, Gestalt and Omelet. The Comprehension and Application questions caused the fifth factor, Mathematics Concepts, to emerge for males. The General Mathematics factor for females was determined primarily by all three problem-solving measures and Mathematics Computation.

Another factor, Numerical Ability, could be considered a mathematics factor for females. However, none of the problem-solving measures had significant loadings for this factor, which was caused by Number Series and Seeing Trends.

One explanation for the sex difference in the number of comparable common factors determined by the three problem-solving scores on the Romberg-Wearne test was that females approached problem solving more systematically. Their methods on the Problem-solving questions paralleled their approaches to the Application questions. Males may have used algorithms and school achievement for the Application questions, but used more of a Gestalt approach for the Problem-solving questions.

Summary

Generalizability of the results of this study was limited by the nonrandom sample, the battery of reference tests, and the difficulty of the Problem-solving questions. The Problem-solving mean (Part III of the Romberg-Wearne test) was only 3.47 and the standard deviation was 2.40. The range of correct responses was 0-13. Fewer than 8% of the sample had over 6 of the 19 Problem-solving questions correct.

Almost all reference tests were selected from a battery used for the CAA Project (Harris & Harris, 1973). The investigator attempted to select from these "concept attainment" tests those she believed to be related to problem solving. The selected battery accounted for 44% of the variance of the Problem-solving questions, 62% of the variance of the Application questions, and 42% of the variance of the Comprehension questions. Significantly, the variance of the mathematics concepts tests of the CAA study, accounted for by the complete battery of reference tests, ranged from .39 to .61.

The Problem-solving questions of the present study appeared to be highly related to the "concept attainment tasks" as were some of the mathematics concepts tests of the CAA study. The Application questions were more highly related to concept attainment than were any of the mathematics concepts tests of the CAA study.

In conclusion, the study suggested the following:

1. Intellectual structures contain a factor specific to mathematics.

2. Problem Solving appears to be related to Numerical Ability.

3. Prerequisite mathematics skills and concepts are related to, and account for, some of the variance of problem solving. However, knowing these skills and concepts does not guarantee successful problem solving.

Chapter 10

Sex, Visual Spatial Abilities, and Problem Solving

Ann Schonberger

This study was initiated in 1975, International Women's Year, during which time the attention of the world was drawn to women's struggle for equal participation in all society's activities. Important to such equal participation is the ability to solve mathematical problems. In the United States men far outnumber women in occupations requiring high mathematical competence. Only hypotheses for causes of this imbalance exist. Some possible reasons are sex bias in career counseling, discrimination in admission to specialized schools, and differences in sex-role socialization. In addition, inherent differences in mathematical ability have been suggested (Carnegie Commission on Higher Education, 1973). Some have said that while girls may be more proficient in computation, boys excel at mathematical reasoning (Jarvis, 1964; Maccoby, 1966). If this is true, mathematical reasoning could be the "critical filter" (Sells, 1973) in the scientific and technical job market, since in these occupations the application of mathematics to problems is valued more highly than computational proficiency.

Employment is not the only area in which women's equal participation will depend on their ability to solve mathematical problems. As consumers of housing and transportation as well as food and clothing, women will need to solve practical mathematical problems. Women should also share equally in the intellectual activities of society:

> Solving problems is the specific achievement of intelligence, and intelligence is the specific gift of mankind: solving problems can be regarded as the most characteristically human activity. (Polya, 1962, p. v)

Is it true that males are better solvers of mathematical problems than females? If so what other differences between men and women might be involved? An area of cognitive abilities which might be related is visual spatial abilities. The development of sex-related differences in these abilities parallels or precedes the development of differences in mathematics achievement (Fennema, 1974; Maccoby & Jacklin, 1974). The use of charts, diagrams, and graphs in all branches of mathematics certainly argues for the logic of this connection. Questions about sex-related differences in mathematics and in spatial abilities and the relationships between the two, as well as questions about the role of drawing diagrams as a link between the two abilities, provided the impetus for this study.

To establish the limits of this investigation some of the key terms in the questions were defined. The author used Zalewski's (1974) subject-dependent definition of a mathematical problem.

> A mathematical problem is a statement which meets three conditions:
>
> 1. The statement presents information and an objective whose answer is based on that information;
> 2. The objective or answer can be found by translation of the information into mathematical terms or application of rules from mathematical areas such as arithmetic, algebra, logic, reasoning, geometry, number theory or topology; and
> 3. The individual attempting to answer the question or attain the objectives does not possess a memorized answer or an immediate procedure. (pp. 4-5)

The third part of this definition serves to differentiate real problems from exercises, but introduces a difficulty in that a problem for one person may be merely an exercise for another.

> Mathematical problem solving is the process of attaining the objective specified in a mathematical problem.
>
> Mathematical problem-solving ability is the ability measured by a test of mathematical problems.

The last definition, a functional one in terms of a test score, makes no assumptions about the components or origins of this ability.

In order to define the visual devices for this study, a system for categorizing external representations of problems was necessary. Incorporating ideas from other category systems (Bruner, 1964; Heimer & Lottes, 1973), the following definition was used.

> A pictorial representation has physical characteristics that can be viewed, but not felt or manipulated independently of the medium in which it is presented. Pictorial representations of objects usually disregard some of the objects' attributes.

The following definitions are related to the mode of representation of a mathematical or spatial problem.

> A diagram is a pictorial representation of information presented in a problem or deduced from information in the problem.
>
> Visual spatial abilities are those measured by tests recognized in the field of cognitive abilities as spatial, whose stimuli are pictorial representations. According to Werdelin, an aspect common to all such tests is "the

ability to comprehend the visual organization of the material and reorganize it" (Werdelin, 1961, p. 77).

A two-dimensional test of visual spatial ability is a test in which the stimuli are planar geometric figures or pictorial representations in which one dimension has been ignored.

A three-dimensional test of visual spatial ability is one in which the stimuli are pictorial representations in which all three dimensions have been drawn in perspective.

Background

To clarify the issues involved in this study, the literature on visual spatial abilities, mathematical problem-solving ability, and the relationships between the two types of abilities was reviewed.

Visual Spatial Abilities

The literature on spatial abilities deals with several questions relevant to this study. Is visual spatial ability a unitary trait or a cluster of several abilities? If there are several, how should the factors (called *spatial factors* or *factors* in this chapter) be described and what tests define them? Are there sex-related differences in the structure of the spatial factors or in performance on spatial ability tests? Do all researchers agree on what tests of spatial ability are?

Factor analyses of spatial ability data gathered during World War II and afterward suggested two or three subfactors of visual spatial ability. Michael, Guilford, Fruchter, and Zimmerman (1957) synthesized previous research considering complexity of stimuli, amount of manipulation involved, movement of parts versus movement of the whole, the subject's body orientation, and the relative importance of speed and power. Their synthesis was influential. The authors were active in writing the factor descriptions and selecting tests for the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, & Price, 1969a, 1969b) developed under the auspices of the Educational Testing Service and referred to in this chapter as the ETS Kit. The spatial factors described in the ETS Kit are basically those of the Michael et al. synthesis, and a number of later studies, such as the National Longitudinal Study of Mathematical Abilities (NLSMA) (Romberg & Wilson, 1969), followed that framework. For these reasons it will be described in detail.

The first factor, Spatial Relations and Orientation (SR-O), was described as the ability to comprehend the arrangement of elements within a visual stimulus pattern with the subject's body as a frame of reference. In SR-O tests parts of the figure remain related to each other in the same way, as the figure as a whole is moved into a different position. The items are usually quite easy and speed is often important.

The second factor was called Visualization (Vz). On Vz tests the subject is expected to mentally manipulate one or more objects or parts of a configuration according to explicit directions. The subject must then recognize or draw the new configuration. Stimuli are generally more complex in Vz tests, and speed is usually less important. The crucial difference between Vz and SR-O tests, according to Michael et al., is that in SR-O tests the figure as a *whole* is rigidly transformed, whereas in Vz tests the figure is broken up into *parts* and the parts are transformed. Kinesthetic Imagery (K), a third factor, appeared to involve right-left discrimination; tests of the K factor have not shown a relationship to mathematical problem-solving ability and will not be discussed further.

Guilford (1967) located these spatial factors in his three-dimensional structure-of-intellect model, in which each cell represents a factor or ability described by an operation on certain content with a specified product. The SR-O factor was labeled Cognition of Figural Systems-Visual, and Vz was called Cognition of Figural Transformations. According to Guilford, SR-O and Vz tests differ only on the product dimension. Burt's (1949) review of factor analytic studies further divided the spatial factor into two- and three-dimensional categories.

The SR-O, Vz framework is not without problems, however, and factor analytic studies have not always shown both factors. In a study of Swedish high school males Werdelin (1958) investigated both the SR-O, Vz division and the dimensional division. Factor analysis yielded only one spatial factor with but a faint indication of an SR-O, Vz subdivisi . In Werdelin's subsequent study (1961) of high school males and femal nalysis of the males' data alone yielded both an SR-O and a Vz factor, but analysis of the combined data turned up only one spatial factor. Separate analysis of the females' data was not reported. Other studies have indicated different spatial factor structures for males and females, although not necessarily identifying SR-O and Vz factors (Harris & Harris, 1973; Very, 1967). French (1965) did not find the SR-O, Vz subdivision even in an all-male sample. However, his study indicated that subjects were using different methods to solve spatial items, some visual and some logical or analytical. Barrett (1953) had also reported different styles of solving items on five different spatial tests.

In addition to the sex-related differences in the factor structure of visual spatial abilities, sex-related differences in mean performance on spatial tests have been noted in a number of reviews (Fruchter, 1954; Garai & Scheinfeld, 1968; Maccoby, 1966; Smith, 1964; Tyler, 1965). To understand these differences better, one should know at what ages and with what kinds of tests they have been observed. Studies of preadolescents published since 1965 show few sex-related differences (Anglin, Meyer, & Wheeler, 1975; Maccoby & Jacklin, 1974), but some have appeared (Harris & Harris, 1973) on two-dimensional SR-O, Vz tests. Sex-related differences become more apparent in

adolescent groups. Three large-scale studies found males' performance superior to females' on a three-dimensional Vz test of the surface development type (Bennett, Seashore, & Wesman, 1973; Droege, 1967; Flanagan, Davis, Dailey, Shaycroft, Orr, Goldberg, & Neyman, 1964). Others have reported significantly higher means for males on tests resembling three-dimensional Vz tests (Bock & Kolakowski, 1973; Stafford, 1961). There is also some evidence of higher male performance on two-dimensional SR-O tests (Flanagan et al., 1964; Hobson, 1947; Thurstone, 1958). Two studies using a variety of spatial tests with college subjects reported sex-related differences in favor of males (Sherman, 1974; Very, 1967).

The last question posed about visual spatial abilities was whether or not researchers agree as to what tests are visual and spatial. Some tests which do not fit into the SR-O, Vz classification involve aural or tactile perception, mechanical knowledge, or motor skills. Since these have shown little relationship to the ability to solve mathematical problems they were disregarded in this study. On the other hand, there is a group of tests whose classification as spatial has not always been recognized, but whose relationship to mathematical problem solving has often been observed. These tests, which require the subject to pick a simple figure out of a more complex stimulus pattern, have been called Gottschaldt's Figures, Concealed Figures, Embedded Figures, Hidden Figures, and Hidden Patterns (see Thurstone & Jeffrey, 1956). The corresponding ability has been named Gestalt Flexibility, Flexibility of Closure, or Convergent Production of Figural Transformations. It has been regarded as a cognitive *style* rather than an ability and labeled Field Independent-Dependence.

That these tests should be classified as spatial was argued by Sherman (1967) and supported by her own research (1974). Maccoby and Jacklin (1974) followed Sherman in considering the tests spatial, a change from Maccobys (1966) categorizing them as measures of field independence. Other evidence of a spatial component comes from the ETS Kit manual (French et al., 1969a, 1969b), from French's (1965) study, and from Guilford's synthesis (1967). The nature and size of the spatial component as well as the identity of its other components is unresolved at present. Also, there appear to be different styles of solving items of this type (French, 1965). For these reasons this author prefers to call them tests of visual disembedding, a descriptive term that makes the fewest assumptions about the underlying cognitive processes.

The demonstrated relationship of these tests to mathematical problem solving, as well as the significantly better performance by males on such tests, were the reasons for this study's concern with tests of visual disembedding. Sex-related differences have appeared in a number of studies summarized by Witkin, Dyke, Faterson, Goodenough, and Karp (1962) and have been the subject of a great deal of controversy, although the differences are generally of small magnitude (Kagan & Kagan, 1970). In fewer than half of the more

recent studies reviewed by Maccoby and Jacklin (1974) did sex-related differences appear; there was some indication that such differences paralleled those in other spatial tests both in magnitude and in time of appearance.

In summary, it appears that spatial tests can be categorized as SR-O or Vz tests with additional subdivision based on dimension. Tests of visual disembedding can also be considered spatial although this has not always been accepted. There is some indication that the structure of the spatial factor is different for males and females. Sex-related differences in performance in favor of males, appearing in adolescence, have been found in a number of studies, especially on three-dimensional Vz tests. Even the largest of these differences in means are usually less than half a standard deviation so the within-sex variation is definitely greater than the between-sex variation.

Mathematical Problem-solving Ability

One might consider items from most types of spatial tests to be problems in transformational geometry. Are women similarly handicapped in solving all other types of mathematical problems? If so, at what age do males begin to outperform women in mathematical problem solving? To answer the latter question studies were grouped as follows: elementary, grades 7 and 8, grades 9 through 12, college, and adult.

To qualify for inclusion in this review a study must have used test items intended to measure mathematical abilities other than computation and test items which seemed to this reviewer to satisfy the definition of a mathematical problem given in the introduction. For example, studies of mathematical reasoning were often included. Selecting studies involved judgment because of the definition's requirement that the individual attempting to answer the question must not possess a memorized answer or an immediate procedure. The general policy was to include doubtful studies.

Several of the studies which contributed the most to this review are longitudinal and a word needs to be said about their methodology. One is the National Longitudinal Study of Mathematical Abilities (NLSMA) (Romberg & Wilson, 1969), probably the most intensive and extensive study in this area. Three different groups were tested: one in grades 4 through 8, another in grades 7 through 11, and the third in grades 10 through 12. A content (number systems, geometry, algebra) by level of behavior (computation, comprehension, application, analysis) matrix was used to categorize the mathematics scales. Both the application and analysis scales are included in this review, but the definition of analysis items seems closer to this study's definition of a mathematical problem. The NLSMA study, designed to compare certain textbook series, involved primarily college-capable students. Another factor which should be considered in evaluating the results is that the sex-related differences reported were those remaining after removal of the variance due to verbal and nonverbal IQ and mathematics achievement. A second longitudinal study (Hilton & Berglünd, 1974), whose results are reviewed, mea-

sured the same students in grades 5, 7, 9, and 11 using the Sequential Test of Educational Progress-Mathematics (STEP-Math) (Cooperative Test Division, 1956-72) which those authors regarded as a measure of the ability to apply skills to problem solving. The sample was divided into an academic group and a nonacademic group according to what program they eventually pursued in high school, and results were analyzed by group.

In the NLSMA study of grades 4 through 6, boys excelled on two out of three application scales, both concerned with number systems, and on the only analysis scale, a geometry scale (Carry & Weaver, 1969). Hilton and Berglünd (1974) found no significant differences between girls and boys in either group on STEP-Math. In a study using fifth-grade subjects (Harris & Harris, 1973), no sex-related differences were found on either of two cognitive abilities tests containing mathematical problems. Similarly, no differences between boys' and girls' performance on an arithmetic reasoning test were found by Parsley, Powell, O'Connor, and Deutsch (1963). A second study (1964) by Parsley, Powell, and O'Connor indicated better performance by males in 12 subgroups and by females in seven subgroups out of a total of 75 comparisons. In a study of sixth-grade students Jarvis (1964) found that boys of all ability levels surpassed girls in arithmetic reasoning. Clearly, although some differences have begun to appear in upper elementary school, the results are mixed.

Sex-related differences were more apparent in the studies reviewed using seventh- and eighth-grade students. Hilton and Berglünd (1974) reported a difference in favor of boys on STEP-Math in the academic group. The NLSMA also gave STEP-Math to one group in seventh grade, categorizing it as an application test, and found boys' performance to be superior (McLeod & Kilpatrick, 1969). Sex-related differences in favor of boys were also found on all but one of the analysis scales and on the one application scale designed by NLSMA (Carry, 1970; McLeod & Kilpatrick, 1969). The content of the scales on which differences were found was number systems and geometry; the scale on which none were found was an algebra scale. In a study of problem-solving styles in high-ability, eighth-grade subjects Kilpatrick (1967) found that although scores for boys and girls were about the same, girls used significantly more deduction and more equations. In the National Assessment of Educational Progress (NAEP) consumer math skills were measured by a test of problems given to 13-year-olds, 17-year-olds, and young adults, ages 26 to 35. In the youngest group the boys' median was one and one-half percent above the median of the total group and the girls' median was one and one-half percent below (Ahmann, 1975).

With the exception of the NAEP all the studies discussed in this section in which sex-related differences were observed were conducted with students of above-average ability. There is another indication that overall superiority of boys in mathematical problem solving in grades 7 and 8 may be due to

superior performance by boys of high ability. In a study of mathematical precocity Stanley, Keating, and Fox (1974) found that in a sample of seventh- and eighth-grade students who volunteered for screening with the Scholastic Aptitude Test-Quantitative (SAT-Q) boys far outperformed girls and the discrepancies increased with age.

Surveying the studies of high school students required additional caution because required mathematics courses are often tracked and mathematics becomes elective in the upper grades. Good examples of this lack of control for number or type of mathematics courses taken are the Project Talent Survey (Flanagan et al., 1964) and the NAEP (Ahmann, 1975) both of which found sex-related differences in favor of males. In all the high school studies reviewed here the students were in the same class or track when tested.

Information on sex-related differences in the NLSMA were reported only for the college preparatory group. At the applications level boys in grades 9 through 11 excelled over girls on five of 12 geometry scales and one algebra scale. At the analysis level the boys' performance was superior on half the algebra and number systems scales; on the geometry analysis scales boys excelled on six of the eight and girls on two (Kilpatrick & McLeod, 1971a, 1971b; McLeod & Kilpatrick, 1971; Wilson, 1972a, 1972b). The impression of overwhelming evidence of male superiority on NLSMA mathematical problem-solving tests should be tempered by several limitations of the study. The restriction to upper-ability students was more severe in the high school data than in the junior high data. The statistical removal of variance due to verbal and nonverbal IQ and mathematics achievement may have left only a small fraction of the variance. Application of the ω statistic (Hays, 1973) to three of the analysis scales given in grade 11 showed that on each, less than one percent of the variance was due to sex. Sex-related differences in performance on the analysis scales appeared most pronounced in the area of geometry, which may be related to males' advantage on spatial items. One of the two geometry scales on which girls excelled was Structure of Proof, which appeared to require verbal rather than spatial skills. Finally, the content of the number systems problems for grades 4 through 11 should be considered. Among these were all the problems about people. In virtually all cases in which sex of a person was specified, the person was male.

Evidence of the importance of these issues was found in other high school studies. In the Hilton and Berglünd (1974) study boys from the academic group scored significantly higher on STEP-Math in grades 9 and 11, whereas in the nonacademic groups boys scored higher only in the eleventh grade. In a study of problem solving in ninth-grade algebra Sheehan (1968) changed a slight (but nonsignificant) advantage of girls into a significant difference in favor of boys by statistically removing variance due to algebra aptitude and previous mathematics achievement and knowledge of algebra. In his

study of high-ability high school students Werdelin (1961) found sex-related differences limited to two tests of geometrical problems.

Studies of college students and adults are even more open to criticism for lack of control for previous exposure to mathematics. Very's (1967) study and the NAEP, both of which found males to be better problem solvers, can be criticized on this point. The most significant body of research on mathematical problem solving in college students is a group of interrelated studies done first at Stanford and then at Yale. After Sweeney's (1953) study which found sex-related differences in addition to those due to intellectual factors, subsequent studies investigated various noncognitive sources of the difference. Carey (1955) found attitude toward problem solving to be a significant factor in males' better performance on the problem test. Following a treatment designed to improve attitude, women's problem-solving performance improved significantly, whereas men's did not. Berry (1958, 1959) and Milton (1957, 1958) investigated the relationship between the Terman-Miles masculinity-femininity index and ability to solve mathematics problems similar to those used by Carey (1955) and Sweeney (1953). In only one of the four studies was the correlation significant after removing effects due to verbal and quantitative factors. In the 1959 study Berry used a number of other noncognitive measures and found that the only ones contributing to the remaining problem-solving variance were two tests of visual disembedding and Carey's attitude test — and this only for males. Milton investigated the effects of problem content and found men superior at solving "masculine" but not "feminine" problems. (Needless to say, the sex-role stereotyping was incredible.) Hoffman and Maier (1966) also investigated the area of problem content but found no significant sex differences.

Summarizing the research on sex-related differences in solving mathematical problems is difficult. As was the case with visual spatial abilities the differences may be small, but they do seem to exist, even after controlling for mathematics background. As with spatial abilities the differences seem to appear in early adolescence and may increase with age until maturity. The studies reviewed in this section indicate that the sex-related differences may be limited to the upper-ability level and to problems whose content is spatial or sex biased.

Relationships between Solving Mathematical and Spatial Problems

The fact that sex-related differences in both visual spatial abilities and mathematical problem-solving ability begin to appear in the upper elementary grades and develop throughout adolescence suggests that there is some relationship between the two abilities (Fennema, 1974). Anglin, Meyer, and Wheeler (1975) and Smith (1964) hypothesized that the importance of spatial ability increases with the cognitive complexity of the mathematical task. In this section the relationship between the two abilities, as seen in several studies, is reviewed with these questions in mind:

1. What is the evidence of a relationship between mathematical problem solving and visual spatial abilities?
2. Is the relationship different for males and for females?
3. If more than one measure of spatial skills was given, are the relationships with the test of mathematical problem solving different for different spatial tests?
4. If more than one problem-solving measure was used, are the relationships with the spatial tests different for different problem-solving measures?

To examine the author's hunch that the common element of spatial ability and mathematical problem-solving tests is actually figural reasoning, comparisons with these tests were made whenever the data were available. As in previous sections, the discussion is organized by grade level.

In the upper elementary grades the CAA Project (Harris & Harris, 1973) provided information on the relationship among spatial abilities, figural reasoning, and mathematical problem solving. In both years of the study, all the correlation coefficients between pairs of these tests were significant. The spatial tests and the mathematical problems tests were more closely related to the figural reasoning tests than to each other, which suggests that figural reasoning may be a bridge between the two. Only for the group in which boys had outscored girls on the spatial test were there any sex-related differences in the correlations. In that group the spatial test was more closely related to the mathematical problems test for girls than for boys. In both this study and the Anglin et al. (1975) study Vz tests were more closely related to mathematical problem solving than were SR-O tests.

With the NLSMA data for grades 5 through 11 there are two ways to investigate the relationship between the spatial tests and the analysis or application measures: one involves correlation coefficients and the other involves analysis of variance. The correlations were made among tests which had been given in different years and were generally low. Despite the probability that none of the differences between correlation coefficients were statistically significant, some interesting patterns can be observed. A two-dimensional Vz test was a better predictor of all the analysis scales than either a two-dimensional SR-O test or a test of visual disembedding. At each grade level the spatial tests were more highly related to the geometry scales than to the algebra or number systems scales. The correlation coefficients decreased with age, probably because mathematical problem solving at upper levels requires more specific mathematical knowledge. The correlations of other mathematics scales in the NLSMA with the spatial tests were no higher and usually lower than those of the analysis scales in all but a few isolated cases (Wilson & Begle, 1972b).

Supporting information on the relationships between the analysis or application scales and the spatial tests was generated by two-way analyses of variance done separately by sex for each pair of tests (Wilson & Begle, 1972a). Significant main effects of the spatial variables on the mathematics measures were found more often for Vz tests than SR-O tests and on geometry scales more often than number series or algebra scales. There appeared to be no pattern in the differences for males and females. Dodson (1972) used a discriminant analysis of a subset of the eleventh grade NLSMA data to characterize successful problem solvers. A test of visual disembedding discriminated among levels of the total problem test and the geometry and number systems subtests; it was not related to the algebra subscale. A two-dimensional SR-O test discriminated among levels of the total test and the geometry subtest and less significantly among levels of the other two subtests.

Werdelin's (1958, 1961) factor analytic studies also demonstrated the close relationship between spatial and problem-solving abilities. Mathematical problem-solving tests, especially geometry or number series tests, loaded on the spatial factors. Two-dimensional Vz tests and three-dimensional spatial tests of both types loaded on Reasoning factors which included all the mathematical problems tests. In the Project Talent study (Flanagan et al., 1964) the three-dimensional Vz tests accounted for more of the variance on each mathematics test than did the two-dimensional SR-O test. There appeared to be no significant sex-related differences in the relationships. As in the CAA Project study (Harris & Harris, 1973) the figural reasoning test was related more closely to both the spatial tests and the mathematical problems test than the two were to each other.

The studies by Berry (1958, 1959) and Sweeney (1953) give some information on the relationship between spatial and mathematical abilities in college students. Sweeney found that matching on performance on a two-dimensional SR-O test was as effective as matching on years of mathematics taken in removing or reducing sex-related differences in performance on his problem-solving tests. Berry found a test of visual disembedding was almost as closely related to his tests of mathematical problem solving as was the SAT-Q.

Thus, in these correlational studies, visual spatial abilities appeared to account, at least as much as any other type of cognitive ability, for part of the variance in mathematical problem solving. The one exception occurred when tests of figural reasoning were included; these tests were more closely related to both spatial and problem-solving tests than the latter two were to each other. Spatial tests were related to problems with different content in this decreasing order: geometry; practical situations or arithmetic; algebra. Tests of the Vz factor were more closely related to mathematical problem solving than SR-O tests; tests of visual disembedding may fall somewhere in between. Whether or not there are sex-related differences in the relationships is unclear.

Correlational studies do not give evidence of cause and effect, but it is usually assumed in the literature that spatial ability is somehow more fundamental than the ability to solve mathematical problems, which involves other components as well. To investigate how spatial skills are used to solve mathematical problems one has to turn to introspective as well as experimental research. Hadamard's (1954) account of his own and Einstein's thinking suggests that the role played by imagery was to record the relationships or patterns among the elements of the problem and to facilitate combining the elements into new patterns. Poincaré (1929) noted that there were individual differences in the use of visual imagery among mathematicians, irrespective of problem content. Menchinskaya (1946) also described this variation among ordinary people solving problems.

In addition to mental images, another aid to problem solving is diagrams — visual images externalized on paper or chalkboard. Two studies on problem solving in geometry (Sherrill, 1973; Webb & Sherrill, 1974) have shown the importance of a correct diagram. Botsmanova (1960) noted the value of pictures illustrating the mathematical structures of arithmetic problems. Two important skills in using diagrams for problem solving seem to be picking out a simple figure from a complex diagram, or visual dimsembedding (Hadamard, 1954; Yakimanskaya, 1970) and recognizing an element of a problem's diagram as the transformed image of a learned theorem's diagram (Kabanova-Meller, 1970). Also important is the ability to represent the information given in a problem by drawing a diagram. There is some evidence that females do less well than males at drawing diagrams (Boe, 1968; Mitchelmore, 1975).

In summary it seems that visual images are only one method that may be used in solving mathematical problems to record the relationships among elements of the problem. Some problem solvers visually transform these elements into new combinations to arrive at a solution. Others rely more on verbal or mathematical symbols to represent and transform the information logically, sometimes with great success. However, Werdelin (1961) pointed out that people who have both visual and verbal methods available are more likely to solve problems successfully than those with only one method at hand.

Designing and Carrying Out the Study

None of the studies discussed in the review of research examined correlations between problem-solving performance and a full range of visual spatial ability measures. Also, different types of problems were either not considered at all or considered only after the fact. Very little research on use of diagrams was found. Thus, by providing partial answers to the questions posed in the introduction, the review makes it possible to replace them with more specific hypotheses.

H1. Boys and girls do not differ in their ability to solve mathematical problems.

H2. There are no sex-related differences in performance on measures of any of the visual spatial abilities: two- or three-dimensional SR-O or Vz visual disembedding.

H3a. There are significant positive relationships between each of the visual spatial abilities and mathematical problem-solving ability.

H3b. The relationships are stronger for Vz tests than for SR-O tests and stronger for three-dimensional tests than for two-dimensional tests.

H3c. These relationships do not differ by sex.

H4. Each type of visual spatial ability is more closely related to solving mathematical problems with high spatial content than those with little spatial content.

H5a. Boys and girls do not differ in their use of diagrams in solving mathematical problems.

H5b. Use of a diagram in solving a mathematical problem is positively related both to the ability to solve that problem and to visual spatial abilities.

H5c. There are no sex-related differences in these relationships.

The review of research indicates that sex-related differences in visual spatial abilities as well as in mathematics achievement begin to appear in grades 6, 7, and 8. For this reason and to avoid the complications caused by different course offerings in mathematics in high school, junior high school students were used as subjects in this study. The entire seventh-grade class of the Fifth Street Junior High School in Bangor, Maine, was selected as the sample. Bangor's population is among the most heterogeneous in the state and its neighborhoods are small enough so that each junior high school encompasses a number of socioeconomic levels. Its population is fairly conservative with respect to sex roles but a citizens' committee had been studying sex roles in the public schools and reporting to the School Committee. Of the three junior high schools in Bangor, Fifth Street was chosen as the most representative by the director of testing and research for the school system because it was always the median. Virtually all of the students in the sample were white and spoke English as a first language. The choice of seventh grade rather than eighth grade was made to avoid complications because some eighth-grade students in this school take two semesters of algebra, some one semester, and some none at all. The seventh-grade mathematics classes were tracked into two levels (Level 1 being the upper one), but essentially the same material was taught in each track. Data were collected in late spring of 1975.

A number of different considerations entered into the choice of tests of visual spatial abilities. They had to be short, easily scored, paper-and-pencil tests. Whenever advisable, tests in related studies were used to facilitate com-

parison of results. To investigate both the SR-O, Vz division and the dimensional division, a two-by-two matrix was constructed and a test chosen for each cell. (See Figure 1.) Both two-dimensional spatial tests, Card Rotations and Form Board 2, were chosen from the ETS Kit tests modified for use by the NLSMA (Wilson Cahen, & Begle, 1968d). The three-dimensional SR-O test was Cubes Comparison, an ETS Kit test (French et al., 1969a, 1969b). The three-dimensional Vz test was the Differential Aptitude Test (DAT) Space Relations (Bennett, Seashore, & Wesman, 1972). The test of visual disembedding chosen was Hidden Figures 2, also an ETS Kit test modified by NLSMA (Wilson et al., 1968d). This is a two-dimensional test; the author knows of no three-dimensional test of this factor.

	SR-O	Vz
two-dimensional	Card Rotations	Form Board 2
three-dimensional	Cubes Comparison	DAT Space Relations

Figure 1. A matrix of spatial tests used in this study.

Choosing a valid and reliable written test of mathematical problem solving for the seventh-grade students was a problem. A decision to base the test on problems from Zalewski's (1974) written-test item bank was made for several reasons. In his study the written test did not quite account for 50% of the variance on the interview test but it came close, so concurrent validity was considerable. Content validity appeared to be at least as high as that of commercial tests. Since this study was designed to include different types of problems as one of its dimensions, selecting problems from a pool was preferable to using an intact test.

In order to investigate the relationships between sex and visual spatial abilities with problems differing in amount of spatial content, three categories were established.

A. Problems in which the stimulus (presentation of the problem) is partly pictorial or which require spatial or geometric skills or knowledge for solution.

B. Problems with a completely verbal stimulus in which spatial skills (such as visualizing the situation or drawing a diagram) may be useful but are not necessary for solution.

C. Problems which appear to have no spatial content. (In other words, any problems that do not fit into categories A or B.)

Four mathematics teachers, two male and two female, classified the entire written test item pool into these categories. A reduced pool was formed of those problems assigned to the same category by all four judges. Some problems were eliminated because they were quite similar to items on the spatial tests or because their content was judged unfamiliar to the subjects of the study by their teachers. Finally, eight problems from each category were chosen from the remaining pool. Since sex-related differences were to be investigated in this study, it seemed appropriate to word the problems to control for sex bias. Whenever possible the problem was made neuter, such as by replacing "boys" or "girls" with "students." Where this was not possible names and pronouns were adjusted so that there were equal numbers of male-acted and female-acted problems in each category.

All these measures were pilot tested with classroom-sized samples at another junior high school in Bangor to see if the tests were appropriate for seventh-grade students and to check the time necessary for administering the tests. Reliabilities for the spatial tests ranged from .47 to .89. The coefficients were not impressive but not much lower than those reported in the literature for some eighth-grade groups. Also the pilot samples were small (16 to 24 students). Some additions were made to the instructions to ensure that the subjects in the main study would understand the tasks. Since the problem test had been newly constructed for this study it was examined item by item after the pilot test. Two items were replaced by different ones.

In the main study data were gathered from 176 subjects in three different testing sessions with makeup tests given a few days later. In the end there were very few missing scores: four subjects missed session 1, none missed session 2, and three missed session 3. The guidance office supplied information on sex of students and IQ in stanines measured by the Otis-Lennon Form J (Otis & Lennon, 1970) given in the fall of the sixth grade. Level of mathematics class was supplied by the subjects. The five mathematics teachers who taught seventh grade were interviewed to provide additional information about the subjects' mathematics programs in the year they were tested. The most important fact provided in these interviews was that three of the four Level 2 (lower) classes had had no geometry that year, although the teachers thought that their students would be familiar with the geometric concepts used in the problem-solving test.

After the data were gathered, the tests were scored. Card Rotations and Cubes Comparison (French et al., 1969a, 1969b) were scored using the number right minus the number wrong formula. For the other three spatial tests [Form Board 2, Hidden Figures 2 (Wilson et al., 1968d), and DAT Space Relations (Bennett, Seashore, & Wesman, 1972)], the score was the number correct. Four types of scales were used for the test of mathematical problem solving. Each of these was applied to the total set of 24 problems and to each of the subtests generated by categorizing the problems A, B, or C. One

scale indicated the number of correct answers in each category and on the total test: Problem Solutions A, Problem Solutions B, Problem Solutions C, and Problem Solutions T. Another scale counted the number of problems for which diagrams were drawn or for which existing diagrams were modified in a rational way: Diagrams A, Diagrams B, Diagrams C, and Diagrams T. A third scale told how many of these diagrams or modifications represented the information in the problem correctly: Correct Representations A, Correct Representations B, Correct Representations C, and Correct Representations T. The last scale was a ratio of Correct Representations compared to Diagrams expressed as a percent: Percent A, Percent B, Percent C, and Percent T.

There were some decisions to be made in the scoring of the diagrams. In certain problems numbers had been used in a nominal sense to identify elements in a group or in a linear order; these were judged to be pictorial representations. Incomplete or erased diagrams were counted. Even after these decisions were made, the scoring was somewhat subjective so agreement with other raters was desirable. Two other doctoral students in mathematics education scored a sample of 25 tests. Intercoder reliability was computed separately for each of these scales: solutions, diagrams, and correct representations. For each of the first two scales 75 pairwise comparisons (3 coders × 25 tests) were made and ratios of agreement (number of agreements ÷ 75) were computed for each problem. Then the agreement ratios were averaged over the 24 problems to give a reliability coefficient for each scale; these were .99 for the solutions scale and .97 for the diagrams scale. The procedure was the same for the correct representations scale except that the divisor for each problem varied according to the number of diagrams observed; the reliability coefficient for this scale was .80. (See Table 1.) Wherever the number of disagreements exceeded 1.00, the tests and scoring sheets were examined to determine the source of disagreement. In some cases the original two coders had made mistakes in following the scoring instructions; in others it was a matter of interpretation which was expected given the nature of this scale. However, on Problem 6 the author realized that she had not recognized two types of correct representations so that problem was rescored on the whole group of 173 tests.

Data Analysis

In designing the study three types of hypotheses were stated: hypotheses about differences in performance, hypotheses about relationships, and hypotheses about differences in relationships. Verifying them required a number of different statistical methods, and some hypotheses were investigated using more than one statistical technique. Although the sample was not randomly selected, it was considered a random sample of a hypothetical population with

Table 1

Agreement over Coders 1, 2, and 3 on Correct Representations

Problem	Number of diagrams observed	Average number of agreements	Average number of disagreements	Average of agreement ratios
1	19	17.00	2.00	.89
2	21	18.00	2.00	.86
3	0			
4	8	8.00	0.00	1.00
5	22	21.33	.67	.97
6	24	20.00	4.00	.83
7	10	10.00	0.00	1.00
8	0			
9	5	4.00	1.00	.80
10	0			
11	0			
12	1	.33	.67	.33
13	9	7.00	2.00	.78
14	1	.33	.67	.33
15	3	3.00	.00	1.00
16	0			
17	0			
18	0			
19	5	4.67	3.33	.58
20	0			
21	0			
22	0			
23	6	4.67	1.33	.78
24	9	9.00	.00	1.00
		Average of agreement ratios over 24 problems		.80

characteristics as described in the previous section, thus justifying the use of statistics based on the assumption of random sampling.

Sex-related Differences in Performance on the Test (Part A)

This section deals with the investigations concerning sex-related differences in performance, specifically H1, H2, and H5a. Additional refinement was possible using data available on the level of mathematics class of each student. Figure 2 indicates the sequence of the data analysis discussed in this section.

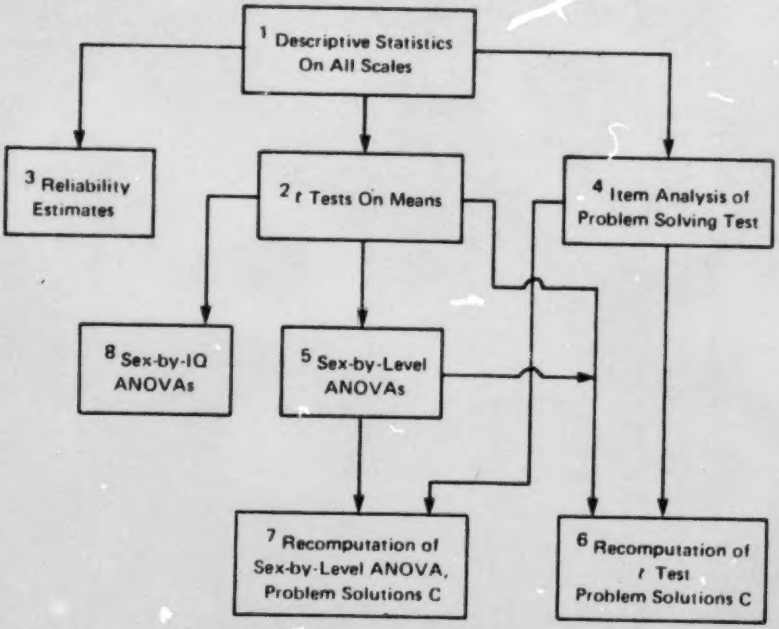

1 Descriptive Statistics On All Scales
3 Reliability Estimates
2 t Tests On Means
4 Item Analysis of Problem Solving Test
8 Sex-by-IQ ANOVAs
5 Sex-by-Level ANOVAs
7 Recomputation of Sex-by-Level ANOVA, Problem Solutions C
6 Recomputation of t Test Problem Solutions C

Figure 2. Sequence of data analysis for Part A.

1 and 2. *Descriptive statistics and t tests on means.* The results of these analyses are reported in Table 2. Of all five spatial tests, only on Form Board 2 was there a significant sex-related difference favoring the boys. Using a ω^2 statistic (Hays, 1973) sex accounted for about 5% of the variance on that test. Significant sex-related differences in favor of girls appeared on six of the problem-solving subscales related to use of a diagram, with sex accounting for 3% to 6% of the variance. There were no significant sex-related differences on the Problem Solutions scales.

3. *Reliability estimates.* The value of these results is dependent on the reliability of the tests. Reliability was estimated for the total group using the Kuder-Richardson Formula 20 (Downie & Heath, 1970) on Form Board 2

Table 2

Descriptive Statistics and Comparison of Means of Boys, Girls, and Both for the Entire Test Battery

Tests	Number of items	Ranges			Means			Standard deviations			Mean Difference *t* value
		Boys	Girls	Both	Boys	Girls	Both	Boys	Girls	Both	
1. Form Board 2[a]	24	17	14	17	6.36	4.90	5.66	3.68	2.92	3.40	3.31**
2. Hidden Figures 2[a]	16	9	10	10	2.42	2.72	2.56	1.76	2.38	2.08	- .93
3. DAT Space Relations[b]	60	41	45	52	26.67	26.07	26.39	9.69	9.48	9.57	.41
4. Problem Solutions A	8	8	7	9	1.67	1.67	1.67	1.45	1.41	1.42	.00
5. Problem Solutions B	8	8	9	9	2.09	2.19	2.14	1.76	1.93	1.84	.35
6. Problem Solutions C	8	9	9	9	3.34	3.00	3.18	1.93	2.02	1.98	1.13
7. Problem Solutions T	24	19	20	20	7.10	6.87	6.99	4.30	4.41	4.35	.35
8. Diagrams A	8	8	9	9	2.78	3.51	3.13	1.49	1.93	1.75	-2.76**
9. Diagrams B	8	5	5	5	1.78	2.16	1.96	1.03	.94	1.00	-2.53*
10. Diagrams C	8	2	2	2	.04	.12	.08	.21	.32	.27	-1.92
11. Diagrams T	24	11	13	13	4.60	5.78	5.17	2.17	2.56	2.43	-3.25**
12. Correct Representations A	8	5	7	7	.89	1.22	1.04	1.03	1.41	1.23	-1.74
13. Correct Representations B	8	4	5	5	.78	1.11	.94	.78	.83	.81	-2.68**
14. Correct Representations C	8	2	2	2	.03	.08	.06	.16	.28	.23	-1.38
15. Correct Representations T	24	6	9	9	1.68	2.37	2.01	1.45	1.98	1.76	-2.59**
16. Percent A	100	101	101	100	28.23	31.26	29.70	29.67	30.32	29.94	- .66
17. Percent B	100	101	101	100	44.06	51.80	47.88	37.73	33.85	35.96	-1.42
18. Percent C	100	101	101	100	60.00	77.78	71.43	54.77	40.10	46.88	-2.35*
19. Percent T	100	101	101	100	34.27	38.77	36.44	25.98	25.87	25.95	-1.14
20. Card Rotations[a]	112	104	104	109	45.57	46.98	46.25	21.44	19.88	20.66	- .44
21. Cubes Comparisons[b]	42	40	50	50	6.67	8.83	7.70	8.69	8.28	8.52	1.43

Note. Number of girls taking all tests = 83.

[a]Number of boys taking these tests = 88.

[b]Number of boys taking these tests = 93. Number of boys taking the remaining tests = 90.

*$p < .05$.

**$p < .01$.

(.75), Hidden Figures 2 (.59), DAT Space Relations (.88), and the four problem-solutions scales of the test of mathematical problem solving (.45, .64, .65, and .80 for scales A, B, C, and T, respectively). A Pearson product-moment correlation coefficient, corrected by the Spearman-Brown formula (Downie & Heath, 1970), was used to calculate the reliability for the Cubes Comparison test (.69) which has two equivalent but separately timed subtests. There was no appropriate method of estimating the reliability of Card Rotations or the scales concerned with diagrams.

4. *Item analysis of the problem-solving test.* This analysis was used mainly to search for sex bias in individual items. Item difficulty was computed separately for boys and girls, as the proportion of correct answers on that item compared to the total number of responses. Sex bias was evaluated for each item by computing a t ratio for the difference between the proportions. One item was found to be significantly ($p < .05$) easier for boys than for girls. This item asks which player has the best shooting record given a table of shots attempted and shots made; boys may have been more familiar with the task than girls. On the whole, however, the test of mathematical problems seemed free of sex bias.

5. *Sex-by-level ANOVAs.* A χ^2 test showed that the actual distribution of boys and girls in the two levels of mathematics classes differed from the expected distribution at the .05 level. (See Table 3.) Since Level 1 students might have had more opportunity than Level 2 students to learn the mathematics needed for the problem-solving test (especially in geometry), and since a disproportionate number of girls was found in Level 1, sex-related differences on the problem-solutions scales or possibly even on the spatial tests might have been obscured. To check this, two-way, sex-by-level ANOVAs were performed on all scores except the percent scales, using only those subjects with complete test data (170 of the original 176). First, a set of exact ANOVAs with equal cell sizes of 32 was computed; the cells were made equal by randomly eliminating subjects from the three cells with more than 32 subjects.

Table 3

Distribution of Boys and Girls in Levels of Mathematics Classes

Level	Boys	Girls
1	43	51
2	50	32

Then, using the same random selection procedures, a second set of ANOVAs with equal cell sizes of 32 was computed. Since there were substantial differences between the two sets of F values, a set of ANOVAs with unequal cell

sizes was computed using the data of all 170 subjects. Table 4 presents the F values generated by these three sets of ANOVAs. Results which were significant on at least two of the ANOVAs for any test were considered important. Others were regarded as artifacts of the samples or due to limitations of this method of analysis.

In general the results of the three ANOVAs were the same as those generated by the t tests: boys performed better on Form Board 2; girls did better on Diagrams A and T. The girls' advantage on Diagrams B disappeared on all ANOVAs as did their superiority on Correct Representations B and T on all but the ANOVA using all the data. In addition, the ANOVAs turned up some interesting sex-by-level interactions on Diagrams T and Correct Representations T. Figure 3, a graph of the cell means on Diagram T for the ANOVA using all the data, shows that the sex-related difference on that scale was due to superior performance by girls from the Level 1 classes. What makes this graph noteworthy is that the graphs of the cell means for *all* the scales involving diagrams or correct representations for *all three* ANOVAs are similar to this one, although few of the interactions are significant at the .05 level.

6 and 7. *Recomputation of the t test and sex-by-level ANOVA: Problem Solutions C.* Another significant ($p < .05$) difference which appeared on

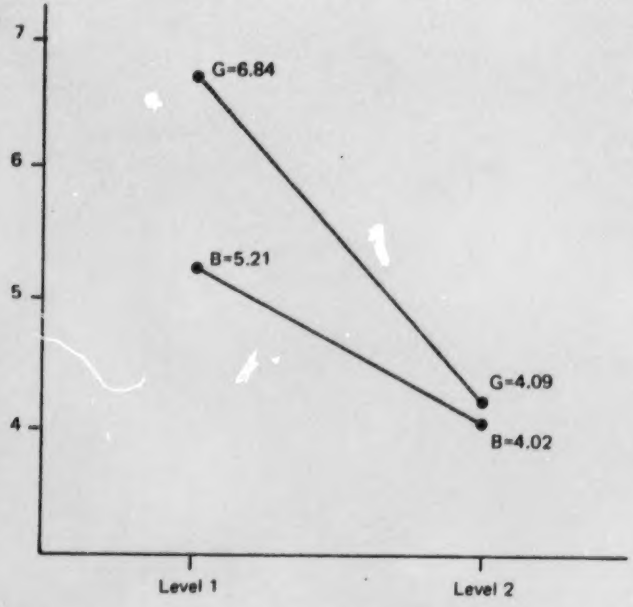


Figure 3. Sex-by-level interaction: cell means on Diagrams T ANOVA using all data.

Table 4

F Values Generated by the Sex-by-level ANOVAs

Tests	Equal cell sizes (1st sample)			Equal cell sizes (2nd sample)			Unequal cell sizes (Using all data)		
	Sex	Level	Sex by level	Sex	Level	Sex by level	Sex	Level	Sex by level
1. Form Board 2	4.63*	2.52	<1	7.49**	<1	6.62*	9.51**	<1	3.01
2. Hidden Figures 2	<1	2.12	2.12	<1	5.08*	4.70*	<1	5.46*	2.80
3. DAT Space Relations	<1	8.45**	<1	<1	6.58*	<1	<1	9.98**	<1
4. Card Rotations	<1	9.33**	1.01	<1	4.00*	<1	<1	5.78*	<1
5. Cubes Comparison	2.20	10.48**	<1	1.10	8.83**	<1	2.48	10.68**	<1
6. Problem Solutions A	<1	16.60**	<1	<1	13.42**	<1	<1	22.91**	<1
7. Problem Solutions B	1.45	39.35**	<1	<1	49.56**	3.56	<1	52.74**	2.10
8. Problem Solutions C	8.77**	60.64**	<1	6.26*	65.22**	<1	6.53*	80.05**	<1
9. Problem Solutions T	5.19*	64.81**	<1	1.41	66.79**	<1	2.78	86.01**	<1
10. Diagrams A	5.17*	37.22**	5.72*	2.75	24.73**	3.14	4.11*	28.88**	3.33
11. Diagrams B	<1	9.52**	2.67	2.05	11.59**	1.57	3.81	13.20**	3.50
12. Diagrams C	2.19	4.93*	<1	1.09	3.02	1.09	1.84	5.33*	1.30
13. Diagrams T	5.09*	37.99**	6.78*	3.85	28.30**	3.86	6.39*	34.32**	5.36*
14. Correct Representations A	<1	18.55**	3.40	1.74	16.88**	5.77*	1.25	22.07**	3.23
15. Correct Representations B	1.20	10.80**	<1	2.59	13.55**	<1	4.97*	16.21**	1.97
16. Correct Representations C	1.87	1.87	<1	<1	<1	<1	1.17	2.34	<1
17. Correct Representations T	1.89	24.25**	3.36	3.51	23.55**	5.36*	3.99*	1.15	4.42*
	$(df_1, df_2) = (1, 124)$			$(df_1, df_2) = (1, 124)$			$(df_1, df_2) = (1, 166)$		

*$p < .05$.

**$p < .10$.

all three ANOVAs was in favor of boys on Problem Solutions C. This difference was also visible, although not significant, on the t test. Since the one problem showing significant sex bias was a C-type problem, the author recomputed some of the statistics for the Problem Solutions C scale eliminating the biased problem. The t ratio was reduced to .65 and the F value in the second sample was reduced to 2.58; both were no longer significant.

Relationships between Visual Spatial Abilities and Mathematical Problem-solving Ability and Sex-related Differences in these Relationships (Part B)

In this section the relationships between the spatial and mathematical variables are examined and the question of whether or not these relationships are the same for both sexes is discussed. Specifically, hypotheses H3a, H3b, H4, H5b, and H5c are investigated. In order to look at the relationships from several points of view a number of statistical analyses were performed. Figure 4 indicates the sequence of these analyses.

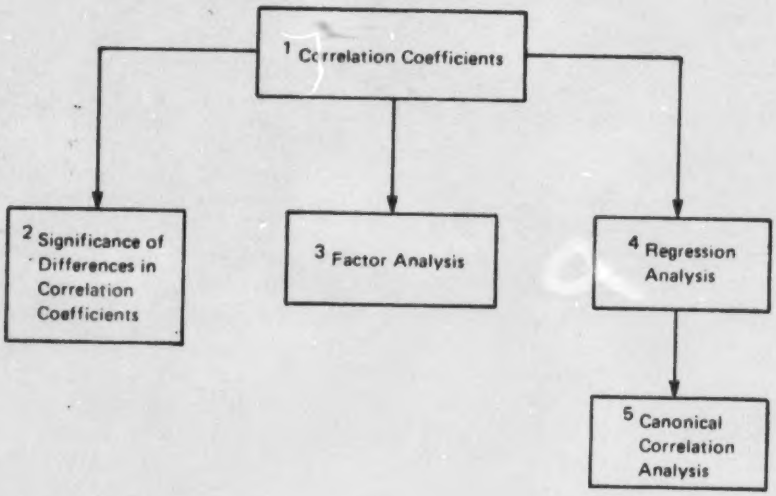

1 Correlation Coefficients
2 Significance of Differences in Correlation Coefficients
3 Factor Analysis
4 Regression Analysis
5 Canonical Correlation Analysis

Figure 4. Sequence of data analysis for Part B.

1. *Correlation coefficients.* Correlations matrices for boys, girls, and the whole group were computed using the spatial scores, the problem-solving subscales, and IQ stanines of the 170 subjects who had taken all of the tests. For girls and for the whole group the relationships between all the spatial tests and all the problem-solving solutions scales were significant and positive. For boys only DAT Space Relations (Bennett et al., 1972) was significantly re-

lated to all the solutions scales. In addition for boys, Cubes Comparison and Card Rotations (French et al., 1969a, 1969b) were significantly related to Problem Solutions B, C, and T, but there were no other significant relationships between the spatial variables and solution scales. For girls and for the whole group the spatial scales, except Form Board 2 (French et al., 1969a, 1969b) were significantly related to Correct Representations A, B, and T. For boys only DAT Space Relations and Card Rotations were significantly related to correctly representing problems with diagrams. The relationships between correctly representing a problem with a diagram and solving that problem are summarized in Table 5. As expected, the size of the correlation coefficients was directly related to the type of problem.

Table 5

Correlation Coefficients Selected to Show the Relative Importance of Using Diagrams in Solving the Three Categories of Problems

	Girls	Boys	Both
	Problem Solutions A		
Diagrams A	.362	.305	.326
Correct Representations A	.702	.662	.669
	Problem Solutions B		
Diagrams B	.359	.109	.234
Correct Representations B	.538	.235	.399
	Problem Solutions C		
Diagrams C	.055	-.011	.107
Correct Representations C	.108	-.066	.028

2. *Significance of differences in correlations coefficients.* It appeared that the correlation coefficients were generally larger for girls than for boys, especially between the spatial and mathematical variables. To investigate the significance of these differences, a Fisher r to Z transformation (Downie & Heath, 1970) was performed on each coefficient and the differences tested for significance using a t ratio. Of the total of 231 pairs tested one could expect that, by chance, 12 would be significantly different at the .05 level. While the actual results (17 at the .05 level) were not much above the chance expectation, the pattern of results is interesting. All but three of the 17 correlations on which boys and girls differed involved drawing diagrams for C-type problems which supposedly have no spatial content. For the boys this was related to Form Board 2; for girls it was related to other diagram drawing scales. The results of this analysis turned up very few differences, but the relationships may contain other sex-related differences that were too slight to detect.

The relative importance of Vz and SR-O tests or two- and three-dimensional tests in predicting mathematical problem-solving ability were also investigated using this method. For the whole group three-dimensional tests of both types were somewhat better predictors of the total problem-solu-

tions scale than two-dimensional tests, but the difference was significant only for the Vz tests. (See Table 6.) The same pattern was observed with all the solutions subscales, although the *p*-values of the differences were not as small as .05.

Table 6

Correlations Between the Spatial Variables and Problem Solutions Total for the Whole Group

	SR-O	Vz	*t* ratio[a] (SR-O-Vz)
2-dimensional	.350	.234	-1.41
3-dimensional	.395	.446	.76
t ratio (3-D-2-D)	.57	2.72*	

[a] *t* ratio for the significance of the difference between correlation coefficients, correlated data (Downie & Heath, 1970).

*$p < .01$.

3. *Factor Analysis.* To reduce the data a series of factor analyses was performed. The method used in the CAA Project (Harris & Harris, 1973) and by Meyer (1976) suggested the possibility of using several different types of factor analysis, each with orthogonal and oblique rotations, and then finding factors common among them. Principal Components Analysis (P Comp A) and Principal Factor Analysis (P Fact A) were the two methods chosen (Evanson, 1975); more factors were extracted by the latter method than the former. Four factors emerged consistently in all four analyses: a Solutions factor, a Space factor, and two Drawing factors, one for A- and B-type problems and the other for C-type problems. (See Tables 7-10.) The tables list variables under a factor if their loading on that factor exceeded .30 for any analysis for either sex.

Several things should be noticed about the Solutions factor. One involves the loadings of Diagrams A and Correct Representations A. Indeed, in P Fact A the Solutions factor split into two subfactors, one for A-type problems and one for the other types. This supports the idea that solving A-type problems is dependent on using diagrams effectively. A second point is that girls' loading of Hidden Figures 2 is higher than the boys'. This could indicate a sex-related difference in problem-solving styles with Hidden Figures items. Perhaps more boys used visual methods and more girls used logical methods. In P Fact A, Hidden Figures 2 had its own factor. The girls' loadings on Problem Solutions B were substantially less than those of the boys.

A single factor emerged in the P Comp A with all the spatial tests except Hidden Figures 2 loading heavily on it. In the P Fact A a subfactor split

Table 7

The Solutions Factor

Variables	Principal components analysis				Principal factor analysis							
	Orthogonal		Oblique		Orthogonal				Oblique			
	Girls	Boys	Girls	Boys	Girls		Boys		Girls		Boys	
1. Problem Solutions A	72[a]	86	70	84	57	43	37	68	49	41	29	62
2. Problem Solutions B	39	71	37	65	38		63	35	37		54	
3. Problem Solutions C	60	62	60	54	59		72		62		72	
4. Hidden Figures 2	72	26	78	32								
5. Diagrams A	40	50	37	26		67		53		17		38
6. Correct Representations A	72	78	65	65	27	75	11	82		42		81

[a]Decimal points omitted.

Table 8

The Space Factor

Variables	Principal components analysis				Principal factor analysis					
	Orthogonal		Oblique		Orthogonal		Oblique			
	Girls	Boys	Girls	Boys	Girls	Boys	Girls		Boys	
1. Cubes Comparison	75[a]	77	75	76	61	61	42	13	61	08
2. DAT Space Relations	69	74	67	71	56	69	35	13	51	21
3. Form Board 2	71	59	71	56	58	52	13	45	19	36
4. Card Rotations	64	71	60	70	55	54	03	33	32	18
5. Problem Solutions B	52	41	41	31	48	35	19	21	20	20
6. Problem Solutions A	38	06			29	14				
7. Problem Solutions C	34	28			27	17				
8. Correct Representation A	30	01			22	13				

[a]Decimal points omitted.

Table 9

The First Drawing Factor (Problem Types A and B)

Variables	Principal components analysis				Principal factor analysis				
	Orthogonal		Oblique		Orthogonal		Oblique		
	Girls	Boys	Girls	Boys	Girls	Boys	Girls		Boys
1. Diagrams A	55[a]	65	46	71	37	53	63	13	48
2. Diagrams B	86	85	87	87	65	76	24	52	72
3. Correct Representations A	34	29	19	41	20	22	39	04	19
4. Correct Representations B	81	87	79	88	69	77	11	60	80
5. Problem Solutions B	46	06			46	10			

[a]Decimal points omitted.

Table 10

The Second Drawing Factor (Problem Type C)

Variables	Principal components analysis				Principal factor analysis			
	Orthogonal		Oblique		Orthogonal		Oblique	
	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys
Diagrams C	91[a]	94	92	95	87	89	86	88
Correct Representations C	93	91	94	92	86	90	85	85
Form Board 2	- 16	36	- 18	36	- 09	25	- 08	37
Hidden Figures 2	- 09	31			- 02	11		

[a]Decimal points omitted.

along dimensional lines was indicated. Problem Solutions B had a substantial loading on the spatial factor for girls and a more modest loading for boys. The First Drawing factor was for A- and B-type problems. In the P Fact A with oblique rotation, the girls' factor split neatly into two subfactors, one for each type. The appearance of Problem Solutions B on this factor for girls suggests that B-type problems were more closely linked to spatial skills for girls than for boys in this sample. The Second Drawing factor was for C-type problems, which appeared to have no spatial content. Form Board 2 had modest loadings on this factor for boys.

4. *Regression Analysis.* To investigate the relative importance of each of the spatial tests in predicting scores on the problem-solving scales, two regression analyses were carried out; in each analysis the data for boys and for girls were treated separately. The first regression analysis used all the spatial tests including Hidden Figures 2; in the second analysis this test was omitted because it had not appeared in the Space factor (see Table 8).

Tables 11 through 14 list the standardized regression coefficients and coefficients of determination for each of the problem-solving regression analyses. In each case the regression equation predicted more of the variance for girls than for boys. The spatial variables were more important predictors of solving B-type problems than A- or C-type problems. For boys DAT Space Relations was generally the only important predictor for all except the C-type problems. For girls all except Form Board 2 and perhaps Card Rotations were important. With the exception of Correct Representations A and T, none of the other regression equations accounted for as much as 20% of the variance, so the tables containing those regression coefficients are not included. Overall, DAT Space Relations was the best predictor of the scales for drawing diagrams. On the scales for correct representation and percent the girls' equations usually had several significant coefficients. The boys' usually had only one: for Correct Representations A and T it was DAT Space Relations, for Correct Representations B it was Card Rotations, and for Correct Representations C it was Form Board 2.

Preliminary comparison of correlation coefficients indicated that the three-dimensional tests were better predictors of the problem-solutions scales than the two dimensional tests. Regression analysis was used to investigate this hypothesis further. Tables 15 and 16 give the coefficients of determination for predicting the problem-solutions scales from the two- and three-dimensional tests separately. Comparison shows that the three-dimensional tests accounted for more variance on each scale than the two-dimensional tests. The differences, which ranged from 3% to 15%, were larger for girls than for boys.

Table 11

Standardized Regression Coefficients
Problem Solutions A

Variables	Regression 1		Regression 2	
	Girls	Boys	Girls	Boys
Hidden Figures 2	.18*	.11		
DAT Space Relations	.19	.23*	.20*	.23*
Cubes Comparison	.23*	-.04	.24**	-.04
Card Rotations	.16	.06	.18	.06
Form Board 2	-.02	.07	.00	.07
Coefficient of determination	.26	.10	.23	.09

*$p < .10$.
**$p < .05$.

Table 12

Standardized Regression Coefficients
Problem Solutions B

Variables	Regression 1		Regression 2	
	Girls	Boys	Girls	Boys
Hidden Figures 2	.18**	.09		
DAT Space Relations	.22**	.32***	.23**	.32***
Cubes Comparison	.28**	.10	.29***	.11
Card Rotations	.17	.17	.20*	.18*
Form Board 2	.04	-.02	.06	-.03
Coefficient of determination	.38	.24	.35	.23

*$p < .10$.
**$p < .05$.
***$p < .01$.

Table 13

Standardized Regression Coefficients
Problem Solutions C

Variables	Regression 1		Regression 2	
	Girls	Boys	Girls	Boys
Hidden Figures 2	.21**	.03		
DAT Space Relations	.35***	.05	.36***	.05
Cubes Comparison	.05	.26**	.05	.26**
Card Rotations	.02	.15	.05	.15
Form Board 2	.05	-.07	.07	-.07
Coefficient of determination	.24	.13	.20	.13

*$p < .10$.
**$p < .05$.
***$p < .01$.

Table 14

Standardized Regression Coefficients
Problem Solutions T

Variables	Regression 1		Regression 2	
	Girls	Boys	Girls	Boys
Hidden Figures 2	.23**	.09		
DAT Space Relations	.32***	.23*	.33***	.23*
Cubes Comparison	.22*	.15	.22**	.15
Card Rotations	.14	.16	.17	.17
Form Board 2	.03	-.02	.06	-.02
Coefficient of determination	.41	.19	.36	.18

*$p < .10$.
**$p < .05$.
***$p < .01$.

Table 15

Standardized Regression Coefficients Two- and Three-dimensional Comparison Problem Solutions A and B

Variables	Problem Solutions A			Problem Solutions B		
	Girls	Boys	Both	Girls	Boys	Both
Two-dimensional						
Form Board 2	.12	.13	.12	.21	.08	.11
Card Rotations	.28	.15	.22	.32	.33	.33
Coefficient of determination	.12	.05	.07	.19	.13	.15
Three-dimensional						
DAT Space Relations	.24	.28	.27	.29	.37	.33
Cubes Comparison	.28	-.01	.19	.34	.14	.24
Coefficient of determination	.20	.08	.14	.30	.20	.25

Table 16

Standardized Regression Coefficients Two- and Three-dimensional Comparison Problem Solutions C and T

Variables	Problem Solutions C			Problem Solutions T		
	Girls	Boys	Both	Girls	Boys	Both
Two-dimensional						
Form Board 2	.15	-.02	.09	.21	.07	.10
Card Rotations	.17	.25	.21	.30	.30	.32
Coefficient of determination	.07	.06	.06	.18	.11	.16
Three-dimensional						
DAT Space Relations	.39	.08	.23	.38	.28	.33
Cubes Comparison	.08	.28	.19	.27	.18	.23
Coefficient of determination	.19	.13	.11	.33	.16	.24

5. *Canonical Correlation Analysis.* To obtain a clearer picture of the relationship between the composite for each type of problem-solving scale and the spatial variables, a canonical correlation analysis was run on the boys' and girls' data separately. The spatial tests, except for Hidden Figures 2, were used as the independent variables; the three subscores for each type of problem scale were used as the dependent variables in the four different analyses. The only canonical correlation which accounted for more than five percent of the variance in the dependent variable was the first one, the problem solutions composite. (See Table 17.) This canonical correlation summarizes a number of results of the regression analysis. The canonical correlation accounted for a larger part of the girls' correlation than the boys', and scores for both A- and B-type problems were related to that part of the variance predicted by the

spatial variables. In the boys' correlation only the scores for B-type problems were important. In the girls' correlation three spatial tests shared the prediction; for boys, DAT Space Relations was most important.

Table 17

Canonical Correlation Analysis
Spatial Composite with Problem Solutions Composite

Variables	Standardized coefficients	
	Girls	Boys
Dependent		
Problem Solutions A	.39	.06
Problem Solutions B	.72	.99
Problem Solutions C	.05	-.03
Independent		
Form Board 2	.08	-.04
DAT Space Relations	.41	.67
Card Rotations	.34	.38
Cubes Comparison	.48	.20
Amount of the variance in the dependent variables accounted for by this canonical correlation	13.31%	7.78%

Conclusions

The following is a summary and interpretation of the results of this study in the context of previous and concurrent research. The conclusions are organized in terms of the hypotheses stated in the design section.

H1. *Boys and girls do not differ in their ability to solve mathematical problems.*

The bulk of the literature reviewed indicated that in seventh and eighth grades, boys are better than girls at solving problems, especially geometric problems or practical problems with spatial content. This seemed especially true at the upper ability levels. In this sample, none of the differences in the problem-solutions scales evaluated by the *t* test were significant at the .05 level. Significantly better performance by boys on C-type problems was found in the sex-by-level ANOVAs. The possible better performance by boys on spatially oriented problems suggested by the background review does not seem relevant as these were problems with no apparent spatial content. The *F*-values for the sex-by-level interaction were less than 1.0, and examination of cell means showed that these differences were found at both ability levels, discrediting any hypothesis of superior male performance at higher ability levels. The possible explanation is that the difference is due to sex bias in problem content. When the only problem on which the difficulty was significantly different for boys and girls (a C-type problem about sports) was removed, the sex-related difference in favor of boys on C-type problems was no longer sig-

nificant, lending support to this explanation. When Meyer (1976), in the study reported in Chapter 9, analyzed her data by sex she found no significant sex-related differences on any of the Romberg-Wearne problem-solving scales. A middle school study by Fennema and Sherman (1976) also used the Romberg-Wearne problem-solving scales; there was a sex-related difference in only one of the four geographical areas of the city used in the study. In summary, it appears that sex-related differences in problem solving in mathematics are disappearing.

H2. *There are no sex-related differences in performance on measures of any of the visual spatial abilities: two- or three-dimensional SR-O or Vz or visual disembedding.*

Some of the studies reviewed earlier showed sex-related differences occurring in or before grade seven. Meyer's (1976) separate analysis of her cognitive abilities data indicated that boys performed better than girls ($p < .01$) on Spatial Relations, a form-board type test. Fennema and Sherman (1976) found no significant sex differences on DAT Space Relations in their middle school study, and in an earlier study using the same test Sherman and Fennema (1977) found significant differences in only two of four high schools participating. The results of this study fit with those cited above. Sex-related differences on Form Board 2 in favor of boys were significant at the .01 level, but there were no differences on DAT Space Relations, Hidden Figures 2, Card Rotations, and Cubes Comparison. As on tests of mathematical problems, it appears that differences in spatial ability are not as widespread as earlier studies indicated. However, further research on form-board tests would be useful.

H3a. *There are significant positive relationships between each of the visual spatial abilities and mathematical problem-solving ability.* H3b. *The relationships are stronger for Vz tests than SR-O tests and stronger for three-dimensional tests than for two-dimensional tests.* H3c. *These relationships do not differ by sex.*

These hypotheses were investigated in a number of ways in this study. The correlation coefficients between all the spatial variables and the problem-solutions scales were significant and positive for girls; for boys many but not all of the correlation coefficients were significant. The problem-solutions scales had low but significant loadings on the spatial factor for girls but not for boys in the factor analysis. In the regression analysis the spatial variables predicted more of the variance on the solutions scales for girls than for boys. Hidden Figures 2 had high loadings on the Solutions factor for girls but not for boys and contributed more to the regression equations for girls than for boys. The closer relationship between mathematical problem solving and spatial abilities for girls than for boys was also suggested by Meyer's (1976) separate factor analysis of her data. She found that Spatial Relations was relevant to only one

factor for boys but to two factors, one of which seemed to involve mathematical problem solving, for girls.

It is the author's hunch that these sex-related differences in the relationship between solving mathematical and spatial problems are a result of differences in method or style of solving spatial items. Both visual and logical methods of solving spatial items were reported in the background section, as was the close connection between spatial tests and tests of figural reasoning. The close connection between figural or abstract reasoning and mathematical problem solving has been similarily noted. If girls as a group, more often than boys, solve spatial items logically, then the relationships between spatial and mathematical tests should be closer for girls than for boys. Since problem-solving style was not studied directly here, this remains a hunch.

A hypothesis generated by the results of the NLSMA (Wilson & Begle, 1972a, 1972b) and the Project Talent study (Flanagan et al., 1964) is that Vz tests are better predictors of mathematical problem solving than SR-O tests. Indeed DAT Space Relations, a Vz test, often had the largest coefficient in the regression equation for the problem-solutions scales, especially for boys. Cubes Comparison was the next best predictor, especially for girls. Given the fact that several different methods have been reported for solving Cubes items, it might be that Cubes Comparison was a Vz test for this sample. On the other hand, Form Board 2, also a Vz test, was the least valuable predictor. It seems reasonable that the three-dimensional characteristic was the important factor. This was supported by the results of the regression equation. The three-dimensional character of Cubes Comparison and DAT Space Relations may have forced this seventh-grade group to use logical methods more often than they did on the two-dimensional tests. This too remains a hunch about problem-solving style.

The suggestion was made in the introduction and reinforced in the background section that poorer performance by females on tests of mathematical problem solving might be due to deficiencies in spatial skills. The results of this study argue against that suggestion. The one spatial test on which there was a significant difference in favor of boys, Form Board 2, was the least important in the regression equations for both sexes.

H4. *Each type of visual spatial ability is more closely related to solving problems with high spatial content than those with little spatial content.*

The curious result of this part of the investigation was that the solution of B-type problems was more closely related to the spatial variables than the solution scale for A-type problems, especially for girls. This appeared in the factor analysis where Problems Solutions B loaded substantially on the Space factor and in the regression analysis where the spatial variables predicted more of the variance on Problem Solutions B than on Problem Solutions A and C. One possible explanation for this involves the stimuli for the A- or B-type

problems. Five of the eight A-type problems have pictorial stimuli; by definition none of the B-type problems do. It may be that in solving B-type problems subjects used spatial skills more often to visualize the situation in the problem or to organize the information given than they did in A-type problems where the visual organization was already presented. However, this is purely speculative.

H5a. *Boys and girls do not differ in their use of diagrams in solving mathematical problems.* H5b. *Use of a diagram in solving a mathematical problem is positively related both to the ability to solve that problem and to visual spatial abilities.* H5c. *There are no sex-related differences in these relationships.*

An unexpected result of this study was that Level 1 girls scored higher on the diagrams and correct representations scales than Level 1 boys, while at Level 2 there was no sex-related difference. All the Level 1 classes had studied some geometry but only one of the four Level 2 classes had geometric instruction. It may be that girls, who are supposedly more successful at school, were applying what they had learned more often and more successfully. The fact that this difference was significant for the A-type problems but only a trend for the others supports this explanation.

Both the comparison of correlation coefficients and the factor analysis showed that the relationship between drawing a diagram for a problem, especially a diagram that represents the information correctly, and solving that problem was closer for A-type problems than other types. This was expected because of the way the problems were categorized. What was unexpected was that the spatial variables were better predictors of the solution scales than of the three types of drawing scales except for Correct Representation A. This seems again to indicate that the spatial tests involved a substantial reasoning component for this seventh-grade sample.

Certainly all the questions raised in this study have not received their final answers. Replication with other groups is always desirable. More information on the relationship between visual spatial abilities and mathematical problem solving might have been obtained if some of the spatial tests had not been so difficult for these students. The author noticed while scoring the problem-solving test that some subjects appeared to misunderstand the geometric concepts in the A-type problems although the teachers had affirmed their classes' familiarity with these concepts before testing. A comprehension-application measure like the Romberg-Wearne test (Wearne, 1976) would have been valuable. Although information was gathered on the presence and strength of relationships between problem solving in mathematics and visual spatial abilities, the cognitive processes used in solving these items remain undetermined. Ultimately these solution processes may be researched through subjects' own accounts of their methods of solving spatial and mathematical problems in the thinking aloud procedures described elsewhere in this volume.

Despite the limitations of this study, there are implications for educational practice. Recent studies have shown few, if any, sex-related differences in mathematical problem-solving ability, cautioning educators against the myth of male superiority at mathematical reasoning. Such obsolete sex-biased views might be partly responsible for the small number of women electing advanced mathematics courses in high school or college. The results of this study also show the importance of constructing tests free of sex bias. While the test used in this study was designed to be neutral with respect to male and female actors, the results of the item analysis indicate that problem content should also be considered. Although sex-related differences in spatial skills did not seem to yield sex differences in ability to solve mathematical problems, these two abilities were related. Spatial training in mathematics classes and more specific instruction in drawing and using diagrams should be encouraged, especially in light of the frequency of diagramming found in this study and in others such as those by Meyer and Zalewski reported in this volume.

In conclusion, it appears that the inferior abilities attributed to females in spatial and mathematical areas should be reevaluated, just as the inferiority of women has been challenged in legal, social, and economic matters. This is not to say the equality has been achieved. However, the changes that have occurred suggest that further change is possible.

Chapter 11

Relationships Between Selected Noncognitive Factors and the Problem-solving Performance of Fourth-grade Children

Donald R. Whitaker

Among the variables presumed related to success in problem solving are attitudes, values, interests, appreciations, adjustments, temperament, and personality. Such variables have been termed *noncognitive* to contrast them with the *cognitive* variables of intelligence, aptitude, achievement, and performance. This study investigated the relationships between selected noncognitive factors and the mathematical problem-solving performance of fourth-grade children.

The Nature of the Problem

Clues about the noncognitive factors that influence problem-solving performance may be found by examining factors thought to influence overall mathematics achievement. Students' attitudes toward a school subject are thought to affect their achievement in that subject. Likewise, educators believe that teachers' attitudes toward a subject can influence their students' attitudes and achievement in that subject. Research findings, while sometimes inconsistent and inconclusive, usually show low, positive correlations between student and teacher attitudes toward mathematics and student achievement in mathematics (see Phillips, 1973; Torrance, 1966; Wess, 1970). These findings raise the question of cause and effect. Do teachers' attitudes cause student attitudes, or is the effect perhaps in the other direction?

Since an individual's overall mathematics achievement is a composite of achievement in several areas, attitude toward mathematics may also be a composite of attitudes toward aspects of mathematics such as computation and problem solving. Researchers, however, have tended to use single, global measures of attitude rather than investigating attitude toward only one phase of mathematics (see, for example, Dutton, 1962; Phillips, 1973; Reys & Delon, 1968). The study reported here examined the relationships between both student and teacher problem-solving attitudes and student performance in mathematical problem solving.

Though research findings vary, there is evidence of sex-related differences in mathematics (for example, see chapters by Meyer and Schonberger in this monograph). These findings suggested including sex as a variable in the

present study. Fourth-grade students and teachers were selected as subjects for the study, since some research suggests that attitudes toward mathematics are formed during the intermediate grades (Callahan, 1971; Fedon, 1958; Stright, 1960).

The Analysis of Mathematics Instruction Project at the University of Wisconsin Research and Development Center for Cognitive Learning has developed an elementary mathematics program called *Developing Mathematical Processes* (DMP) (Romberg, Harvey, Moser, & Montgomery, 1974, 1975, 1976). The DMP program is a research-based, activity-oriented approach to teaching and learning mathematics in grades K-6. One of the basic goals of DMP is the development of mathematical problem-solving skills and processes. While a DMP staff member, the author worked with a number of teachers and students in DMP schools and was impressed by the manner in which students attack problems and by the positive attitude both students and teachers seemed to have toward the DMP program (Montgomery & Whitaker, 1975). Therefore, the sample for this study involved students and teachers who had participated in the large-scale field test of DMP. For comparison, a non-DMP sample of students and teachers was included.

Key Terminology Used in the Study

For this study a *problem* is a situation which presents an objective that an individual is motivated to achieve, although no immediate procedures are available to arrive at that objective (Zalewski, 1974, p. 2). The situation in each problem is mathematical in nature. *Problem solving* is the process of analyzing the situation posed in a problem, producing a solution procedure, using that procedure, and achieving a solution to the problem. *Mathematical problem-solving performance* is represented by a score on a mathematical problem-solving test.

As used in this study, attitude is the predisposition of an individual to evaluate some symbol, object, or aspect of his or her world in a favorable or unfavorable manner (Katz, 1967). In particular, attitude toward problem solving is the predisposition of an individual to evaluate factors related to mathematical problem-solving in a relatively favorable or unfavorable manner and is represented by a score on an attitude scale.

The Questions of the Study and Their Significance

The first two questions this investigation was designed to answer pertained to the attitudes of the subjects of the study:

Question 1: Do fourth-grade students have favorable attitudes toward problem solving? (Do differences in attitude exist if stu-

dents are classified by sex or program type: DMP versus non-DMP?)

Question 2: Do fourth-grade teachers have favorable attitudes toward problem solving? (Do differences in attitude exist if teachers are classified by type of program taught: DMP versus non-DMP?)

Educators generally desire that students and teachers hold favorable attitudes toward all phases of the school program, so the findings of the study help to determine if this is the case. Directional relationships between problem-solving attitudes and problem-solving performance were analyzed in the second part of the study.

The problem-solving performance of the participating students was of major importance for several questions of the study. Question 3 deals with that issue:

Question 3: How do fourth-grade students perform on a test of problem-solving performance which provides measures of comprehension, application, and problem solving? (Do differences in problem-solving performance exist when students are classified by sex or by program type: DMP versus non-DMP?)

Most tests of problem-solving performance have provided a single score measuring each student's ability to solve problems, but such scores are inadequate to explain why some students are successful at solving a set of problems and others are not. The Romberg-Wearne Problem-solving Test (Wearne, 1976) used in this study was designed to overcome this inadequacy.

Assessing attitudes toward problem solving is justified if there is reason to suspect that these attitudes are related to performance. The fourth and fifth questions of the study pertain to that relationship.

Question 4: What is the relationship between fourth-grade students' attitudes toward problem solving and their performance in problem solving? (Do differences in this relationship exist if students are classified by sex or by program type: DMP, versus non-DMP?)

Question 5: What is the relationship between fourth-grade teachers' attitudes toward problem solving and their students' performance in problem solving? (Do differences in this relationship exist if students are classified by sex or by program type: DMP versus non-DMP?)

Past studies have not examined the relationship between attitude and performance in a problem solving or any other single phase of the mathematics

curriculum. If problem-solving attitudes and performance are highly related, then research into other specific phases of the curriculum is mandated.

Educators generally believe teacher attitude and effectiveness in a particular subject to be important determinants of student attitudes and performance in that subject (Aiken, 1969). However, research findings pertaining to this belief have not been definitive. The last two questions of the study were directed at this cause-effect relationship.

Question 6: Do fourth-grade teachers' attitudes toward problem solving affect their students' problem-solving performance or is the effect of the opposite nature? (Do differences exist when students are classified by sex?)

Question 7: Do fourth-grade teachers' attitudes toward problem solving affect their students' attitudes toward problem solving or is the effect of the opposite nature? (Do differences exist when students are classified by sex?)

It is reasonable to suspect that students' attitudes and performance might affect teachers' attitudes, instead of the relationship being only in the other direction. It is important, then, to gain information on which source — the teacher or the student — has the greater effect on the other's attitude and performance.

Related Literature

A review of the recent related problem-solving literature is given in Chapter 2 of this monograph. This section of the present chapter includes an overview of recent attitudinal research and summarizes studies having particular significance for this investigation.

The Nature of Attitudes

Most definitions (see Allport, 1967) indicate that attitude is a learned state of readiness, a predisposition to react in a particular way toward certain stimuli. Important to any study is the idea that attitude involves both cognitive and noncognitive components — that is, both beliefs and feelings — and, to some extent, a behavioral component. For example, a student's attitude toward mathematics is a composite of intellectual appreciation coupled with emotional and behavioral reactions to the subject.

In a condensation of theoretical formulations about attitudes, Scott (1968) suggests that the concept has as many as 11 variable properties. Of particular importance to this assessment of attitudes toward mathematics are the properties of direction (Does the individual generally like or dislike mathematics?) and intensity (How strongly does the individual feel about this attitude?).

The Measurement of Attitudes

A number of techniques are available to assess attitudes. Corcoran and Gibb (1961) describe several of the measures of attitudes toward mathematics, including questionnaires, attitude scales, incomplete sentences, projective pictures, essays, observational methods, and interviews. Of these techniques, perhaps the most widely used are the attitude scales. The most popular types of scales are the Thurstone scale (Thurstone, 1928), the semantic differential (Osgood, Suci, & Tannenbaum, 1957), and the Likert scale (Likert, 1932), the type used in the present study. Other but less popular measures include biographical and essay studies (Campbell, 1950) and the monitoring of galvanic skin responses of subjects (Cooper & Pollock, 1959). Still other researchers argue for a multiple-indicator approach (Cook & Selltiz, 1964), which infers attitude from subjects' behavior rather than making direct measurements.

Elementary School Students' Attitudes Toward Mathematics

A number of attempts have been made to establish the relationship between attitude and student achievement in mathematics. Studies by Poffenberger and Norton (1959) and by Shapiro (1962) found low positive correlations between the two. The results of the extensive National Longitudinal Study of Mathematical Abilities (NLSMA) suggested a relatively stable pattern of positive correlations of mathematics attitude scores with both mathematics achievement scales and mathematics grades in each population of the study (Crosswhite, 1972). On the other hand, studies by Anttonen (1969), Cleveland (1962), and Faust (1963) failed to support the belief that there is a positive correlation between attitude and achievement in mathematics. Some research has linked general intelligence with attitude toward mathematics (Crosswhite, 1972; Shapiro, 1962).

Evidence suggests that attitudes toward mathematics may be formed as early as the third grade (Callahan, 1971; Fedon, 1958; Stright, 1960), although these attitudes tend to be more positive than negative in elementary school. Interestingly, there is evidence of a decline from third through sixth grades in the percentage of students who express negative attitudes toward mathematics (Crosswhite, 1972; Stright, 1960).

At the elementary school level attitude toward mathematics and achievement in mathematics are related to a number of personality variables, such as good adjustment, high sense of personal worth, greater sense of responsibility, high social standards, motivation, high academic achievement, and freedom from withdrawal tendencies (Naylor & Gaudry, 1973; Neufeld, 1968; Swafford, 1970). Children with positive attitudes toward mathematics tend to like detailed work, to view themselves as more persevering and self-confident (Aiken, 1972), and to be more "intuitive" than "sensing" in personality type (May, 1972). When attitude scores are used as predictors of achievement in elementary school mathematics, a low but significant positive

correlation is usually found (Evans, 1972; Mastantuono, 1971; Moore, 1972).

Elementary Teachers and Attitudes Toward Mathematics

Unfortunately, many prospective teachers seem to have unfavorable attitudes towards mathematics (Dutton, 1962; Reys & Delon, 1968). However, preservice mathematics content and methods courses for prospective elementary teachers seem to improve attitudes toward mathematics (Gee, 1966; Reys & Delon, 1968; White, 1965; Wickes, 1968).

Attitudes of elementary teachers toward mathematics are generally less positive than those of secondary school mathematics teachers (Wilson, Cahen, & Begle, 1968c). On the other hand, Stright (1960) concluded that a large percentage of elementary teachers enjoy teaching arithmetic and attempt to make the subject interesting. Brown (1962) concluded that experienced teachers had more positive attitudes toward arithmetic and possessed a better understanding of the subject than did less experienced teachers. Todd (1966) found that a state-wide inservice course produced significant changes in teacher attitudes toward arithmetic and in arithmetic understanding.

Teacher Attitude as Related to Student Attitude and Achievement

Educators generally believe that teacher attitude and effectiveness in a particular subject are salient determinants of student attitudes and performance in the subject. Poffenberger (1959) concluded that teachers who tend to affect students' attitudes and achievement positively have a good knowledge of and interest in the subject, a desire to have students understand, and good control of the class.

The relationship between teacher attitude and student achievement in mathematics has been verified more often than has the connection between teacher attitude and student attitude. Torrance (1966) concluded that teacher effectiveness had a positive effect on student attitudes toward teachers, methods, and overall school climate. Phillips (1973) found that teacher attitude for 2 of the past 3 years, especially most-recent teacher attitude, was significantly related to student attitude toward mathematics. On the other hand, studies by Caezza (1970), Van de Walle (1973), and Wess (1970) found no statistically significant relationships between teacher attitudes and either the attitudes or attitude changes of their students.

Attitudes Toward Problem Solving

Several years ago Brownell (1942) observed that favorable student attitudes toward problem solving are a desirable and obtainable educational outcome. More recently, Polya (1965) has stressed the importance of favorable teacher attitudes in helping students acquire problem-solving proficiency. In a publication by the Ontario Institute for Studies in Education (1971) the following observation is made:

> Granted that problem solving is both a desirable and an essential part of school mathematics, it seems a necessary prerequisite for successful development of problem solving skills that both teacher and student have positive attitudes to problems. (p. 35)

Aiken (1970) has called for more intensive investigations of attitudes toward mathematics and has suggested that an individual's attitude toward one aspect of the discipline, such as problem solving, may be entirely different from his or her attitude toward another aspect, such as computation. The following is a review of the work of the few researchers who have investigated problem-solving attitudes.

A problem-solving attitude scale for college students. Though Carey (1958) was interested in general problem solving, rather than mathematical problem solving, her study is important because it represents a first attempt to construct a problem-solving attitude scale. She constructed a reliable instrument with a Likert-type format to measure attitudes toward problem solving. The use of this scale enabled her to conclude that college-age men and women differ in attitudes toward problem solving and that problem-solving performance is positively related to problem-solving attitude.

A Brazilian study of problem-solving attitudes. Lindgren, Silva, Faraco, and DaRocha (1964) studied attitudes toward problem solving as a function of success in arithmetic in Brazilian elementary schools using a 24-item adaptation of the Carey (1958) scale. An arithmetic achievement test, a general intelligence test, and a socioeconomic scale also were administered to fourth-grade students. Favorable problem-solving attitudes were positively and significantly correlated with arithmetic achievement, although the correlations were rather low. Problem-solving attitudes of the students showed near-zero correlations with intelligence test scores and socioeconomic status. Unfortunately, Lindgren et al. did not correlate problem-solving attitudes with performance in problem solving.

A problem-solving inventory for children. Of particular interest to the present study is the work by Covington (1966) who devised instruments to assess problem-solving competency among upper elementary school children. This effort resulted in the development of the *Childhood Attitude Inventory for Problem Solving* (CAPS). This inventory is designed to indicate children's beliefs about the nature of the problem-solving process, attitudes toward certain aspects of problem solving, and degree of self-confidence in dealing with problem-solving tasks. Though CAPS itself does not assess attitudes toward mathematical problem solving, it holds promise as a model for similar instruments related to mathematical problem solving.

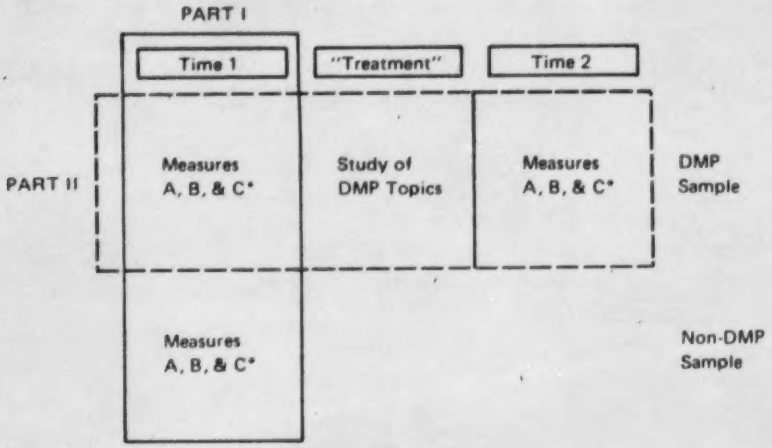
Concluding Remarks

The complex nature of both attitudes and mathematical problem-solving makes the search for definitive answers about the natures of each variable

and their relationships tedious and frustrating. At best, the evidence about the two variables is inconclusive, and research into their relationships is almost nonexistent.

Design and Conduct of the Study

The study was planned to be conducted in two parts with samples from two different populations, as depicted in Figure 1.



*Measure A: Student Mathematical Problem-solving Test
Measure B: Student Mathematical Problem-solving Attitude Scale
Measure C: Teacher Mathematical Problem-solving Attitude Scale

Figure 1. The design of the study.

The following is a description of the instruments and specific details of the design.

The Instruments of the Study

Three instruments were used in the present study: (a) a mathematical problem-solving test, (b) a student mathematical problem-solving attitude scale, and (c) a teacher mathematical problem-solving attitude scale. The mathematical problem-solving test was developed by Romberg and Wearne (Wearne, 1976) and is described in more detail in Chapter 8.

Efforts to develop reliable scales to measure attitude toward problem solving have met with reasonable success. However, existing scales were either inappropriate or unavailable for use in this study. Therefore, two problem-solving attitude scales were developed by the author.

Construction of the student attitude scale. Nunnally (1967) has observed that if verbalized attitude is to be measured, the content validity of the instrument is the major issue. Furthermore, he maintains that content validity is best ensured by a representative collection of items and sensible instrument construction. Therefore, a procedure similar to that used in developing the NLSMA attitude scales (see Romberg & Wilson, 1969) was followed by the author in constructing the student mathematical problem-solving attitude scale. First, a pool of 82 Likert-type items was constructed. Each item was intended to measure some aspect of fourth-grade students' attitudes toward mathematical problem solving. Next, the list of items was submitted to a panel of reviewers for careful scrutiny. Any item rejected by at least two reviewers was discarded. This procedure yielded a 40-item pilot scale with Likert format.

Pilot test and item analysis of the student attitude scale. The pilot version of the student mathematical problem-solving attitude scale was administered by the author to 51 fourth-grade students. Item responses for each student were coded on a five-point scale, ranging from five for the most favorable response to one for the most unfavorable response. Total scale scores could vary from 200 for the most favorable attitude to 40 for the most unfavorable attitude. A score of 120 signified a neutral attitude. Mean total response score was 142.9. Cronbach's alpha (Cronbach, 1951), a measure of internal consistency reliability of the instrument, was .90 for the total scale.

The revised student scale. Following the analysis of the pilot test results, a revised, two-part student mathematical problem-solving scale was developed. Part I had 12 items in a happy/sad faces response format and was designed to provide an informal measure of a student's attitude. An example of the items in this part is shown in Figure 2. Part II consisted of 24 items in modified Likert format providing a formal and more specific measure of attitude. An example of a formal item is shown in Figure 3. The revised scale, as a whole, provides a composite measure of a number of variables which influence a fourth-grade student's attitude toward mathematical problem solving (see Whitaker, 1976).

If we spent more time in school doing math problems, I would be

Figure 2. Example of a mathematical problem-solving attitude item with "happy/sad faces" format.

After I read a problem, I like to think about what I know and what I don't know in the problem.

____ REALLY AGREE

____ AGREE

____ CAN'T DECIDE

____ DISAGREE

____ REALLY DISAGREE

Figure 3. Example of a mathematical problem-solving attitude item with modified Likert format.

Construction of the teacher attitude scale. A procedure nearly identical to that used for the student attitude scale was adopted for construction of the teacher scale. First, a pool of 70 Likert-type items was written, each item to measure some aspect of an elementary teacher's attitude toward mathematical problem solving. Many statements were similar in content and wording to those written for the student scale. The pool was submitted to the same panel of reviewers who examined the student items. Any item rejected by at least two reviewers was again discarded. The same five-part response format — really agree, agree, can't decide, disagree, and really disagree — was used on the teacher scale.

Pilot test and item analysis of the teacher attitude scale. A 50-item pilot version of the teacher mathematical problem-solving attitude scale was administered by the author to 28 elementary school teachers. A five-point coding scheme was adopted for each response so that the maximum possible score on the scale was 250, indicative of a very favorable attitude. A score of 150 indicated a neutral attitude, while a score of 50 meant a very unfavorable attitude. Mean total score for the pilot sample was 181.5. Internal consistency (Cronbach, 1951) of the teacher scale was .96.

The revised teacher attitude scale. After revisions, the teacher mathematical problem-solving attitude scale used in the study consisted of 40 items in Likert format. Thirty-one pilot scale items were used and nine other items were added reflecting the teaching of problem-solving skills and processes. The total scale provides a composite measure of an elementary teacher's attitude toward mathematical problem solving (see Whitaker, 1976).

Part I of the Study

The first part of the study dealt with questions 1-5 formulated earlier in this chapter. The sample and procedures for this part are outlined below.

The sample. Subjects in the sample for Part I of the study were 30 fourth-grade teachers and their fourth-grade mathematics classes. Fifteen of the teachers and students were participants in the large-scale field test of the *Developing Mathematical Processes* (DMP) (Romberg et al., 1974, 1975, 1976) program. They were using the commercial fourth-grade DMP materials during the 1975-76 school year. The 15 DMP classes were in six Wisconsin schools. The remaining 15 fourth-grade classes were in seven other Wisconsin schools. These teachers and students were using mathematics programs other than DMP.

The procedures. During the second week of December 1975, the three instruments of the study were administered by the author and a testing specialist to the DMP students and teachers. Testing was carried out in the classrooms of the participating schools on two different days; the mathematical problem-solving test was given on the first day and the attitude scales on the second day. The non-DMP testing was begun during the second week of January 1976 and completed early in the fourth week of that month. Procedures similar to those used with the DMP sample were followed with the non-DMP sample.

Part II of the Study

The second part of the study dealt with questions 6 and 7 posed earlier in this chapter. The paragraphs below describe the sample and procedures for this part of the study.

The sample. Subjects were the same fourth-grade teachers and their mathematics students in the DMP sample of Part I. Unfortunately, because of a teacher resignation, the second part of the study was conducted with only 14 classes. The non-DMP teachers and students did not participate in Part II.

The procedures. The study involved two different testing periods (Time 1 and Time 2) with an intervening "treament" period. The first testing period was described above. The second testing period began during the second week of March 1976 and ended during the last week of that month. Testing at Time 2 was conducted in the classrooms of the participating schools, again on two different school days. Tests were administered by the author and the testing specialist who assisted at Time 1. The mathematical problem-solving test was administered on the first day. This test was an alternate version of that used at Time 1, except that each item on the second version had a multiple-choice response format. The mathematical problem-solving attitude scales (with rerandomized items) were given the day after the problem-solving test. The intervening "treatment" period between Time 1 and Time 2 lasted approximately 12 weeks, although the duration could not be controlled precisely

because of scheduling difficulties. The "treatment" consisted of instruction in the regular sequence of DMP topics for fourth grade, with the restriction that teachers select at least one topic from the problem-solving strand of DMP. Without exception, teachers elected to cover DMP Topic 57, *The Numbers 0-999,999.*

Findings for Part I

Five main questions were the foci around which the first part of the study was conducted and the data were analyzed. Below is a discussion of the data and findings for each question in turn.

Findings for Question 1

The first question of the study was: Do fourth-grade students have favorable attitudes toward problem solving? To answer this question the 36-item mathematical problem-solving attitude scale was administered to students in the sample. Item responses for each student were coded on a five-point scale ranging from five for the most favorable response to one for the most unfavorable response. A total scale score of 180 represents the most favorable attitude toward problem solving, a score of 108 signifies a neutral attitude, and a score of 36 represents the most unfavorable attitude. The 12 items in Part I of the scale measure students' reactions to general types of mathematics problems, with possible scores ranging from 60 for most favorable to 12 for most unfavorable; a score of 36 represents a neutral attitude. The 24 items in Part II of the scale assess students' reactions to specific problem situations and problem-solving techniques, and scores can range from 120 for most favorable to 24 for most unfavorable, with a score of 72 indicating a neutral attitude.

Table 1 gives a summary of the mathematics problem-solving attitude scores for the 619 students who responded to the scale. Scores ranged from unfavorable to very favorable on each of the two parts of the scale and on the total scale, but each mean score was closer to indicating a favorable attitude than a neutral attitude. Thus, the fourth-grade students in the sample seemed to possess favorable attitudes toward mathematical problem solving. When the data were analyzed by sex and by program type (DMP versus non-DMP), no significant differences in results were observed.

Table 1

Mathematical Problem-solving Attitude Scores of Students in Sample Population (N = 619)

Scale part	Minimum	Maximum	Mean	*SD*
I (Informal)	12.0	60.0	43.7	8.4
II (Formal)	38.0	116.0	85.9	12.9
Total (Composite)	52.0	176.0	129.6	18.9

Cronbach's alpha (Cronbach, 1951) was computed for each part of the scale. For Part I, the reliability coefficient was .85; for Part II it was .82; the total scale reliability coefficient was .88. These reliability estimates were judged to be quite satisfactory.

Findings for Question 2

The second question of the study was: Do fourth-grade teachers have favorable attitudes toward problem solving? To answer this question a teacher mathematical problem-solving attitude scale was administered to the 30 teachers in the sample. Thirty-one of the 40 items on the scale assess teachers' reactions to types of mathematics problems and problem situations, and frustration or anxiety experienced when solving problems. The remaining items assess teachers' feelings about teaching various problem-solving skills and processes. The total scale provides a composite measure of an elementary school teacher's attitude toward mathematical problem solving. Item responses are coded on a five-point scale, ranging from a score of five for the most favorable response to one for the most unfavorable response. A total scale score of 200 represents the most favorable attitude toward problem solving, a score of 120 signifies a neutral attitude, and a score of 40 indicates the most unfavorable attitude.

The attitude scores of the teachers in the sample ranged from slightly favorable to very favorable, evidenced by a minimum recorded score of 134 and maximum recorded score of 175. Mean score for the sample was 156.5 (standard deviation of 9.6), indicating that the teachers possessed favorable attitudes toward mathematical problem solving. When teacher attitudinal data were analyzed by type of mathematics program taught (DMP versus non-DMP), difference in mean attitude scores was not significant.

The internal consistency of the teacher attitude scale was .80. Though the reliability estimate was lower than anticipated, it was judged satisfactory given the relatively small sample size.

Findings for Question 3

The third question investigated was: How do fourth graders perform on a test of problem-solving performance which provides measures of comprehension, application, and problem solving? Three separate scores were reported for each student responding to the mathematical problem-solving test (Wearne, 1976). Students were able to solve correctly more of the comprehension items than application items and more of the application items than problem-solving items; this result was expected since it reflects the order of difficulty of the items. The problem-solving items are the most difficult and are *problems* in the sense of the definition given in Chapter 1. As shown in Table 2, of a total of 22 three-part items on the test, mean number of problems solved correctly by the students was 15.00, 9.50, and 3.19 for comprehension, application, and problem solving, respectively. Most of the students, then, could not be classified as good at solving problems of the type specified by the

definition. A more detailed discussion of student performance on the problem-solving test may be found in Chapter 8.

When the data was grouped by sex, differences in problem-solving performance were not significant. However, when scores were analyzed by program type, DMP students performed significantly better ($p < .01$) than the non-DMP students on the comprehension and application parts of the problem-solving test. Difference in performance for the problem-solving part of the test was not significant.

Table 2

Mathematical Problem-solving Performance Scores of Students ($N = 611$)

Items	Number of items	Minimum/ Maximum	Mean	*S D*
Comprehension	22	2/22	15.00	3.5
Application	22	1/20	9.50	3.9
Problem Solving	22	0/15	3.19	2.5

Findings for Question 4

The fourth question was: What is the relationship between fourth-grade students' attitudes toward problem solving and their performance in problem solving? The correlation matrix calculated to determine this relationship is presented in Table 3. Correlations between the three student attitude scores and the three problem-solving scores are shown. Significant positive correlations ($p < .01$) existed between each of the attitude scores and each of the problem-solving scores. Aside from the strong intercorrelations between the various parts of each instrument, the strongest correlations were found between students' Part II attitude scores and their comprehension, application, and problem-solving scores. When the data were grouped by sex, there was a significant positive relationship ($p < .05$) between attitude and performance for both boys and girls.

Table 3

Correlation Matrix for Students' Mathematical Problem-solving Attitudes and Mathematical Problem-solving Performance ($N = 579$)

	Attitude I	Attitude II	Total attitude	Comprehension	Application	Problem solving
Attitude I	1.00					
Attitude II	.55*	1.00				
Total attitude	.82*	.93*	1.00			
Comprehension	.12*	.24*	.21*	1.00		
Application	.15*	.31*	.27*	.69*	1.00	
Problem solving	.15*	.25*	.23*	.49*	.69*	1.00

*Significant at $p < .01$ as determined by Fisher Z-transformation (see Hays, 1973).

Correlations calculated for the student data categorized by program-type indicated a positive relationship between problem-solving attitude and performance for both groups. Correlations for the DMP sample ranged from .03 to .17, with six of the nine correlations between attitude and performance significant at the .05 level. For the non-DMP sample, the correlations were somewhat higher, ranging from .18 to .43, with all correlations significant at the .05 level. Thus, there appeared to be a stronger relationship between student problem-solving attitude and problem-solving performance for the non-DMP sample than for the DMP sample. Exploratory analyses with data from the DMP sample suggested that students with high problem-solving performance have problem-solving attitudes considerably higher than average, while those students with low performance have lower than average attitudes.

Findings for Question 5

The fifth question investigated was: What is the relationship between fourth-grade teachers' attitudes toward problem solving and their students' performance in problem solving? Correlations between teachers' attitudes and the mean problem-solving performance of the students in their classes were found to be consistently very weak, negative, nonsignificant, and in the range of -.05 to -.08. Thus, for the 30 classes in the sample, there appeared to be little observable relationship between teacher problem-solving attitude and student problem-solving performance. No significant differences were found when the data were analyzed by sex of the students.

Surprising and almost unbelievable results were found when correlations were computed on the basis of program-type. For the non-DMP sample the correlations between teacher attitude and mean student problem-solving performance ranged from .16 to .19 and were nonsignificant. However, for the DMP sample, substantial negative correlations were found; they ranged from -.47 to -.59 and two of the three were significant at the .05 level. In an attempt to explain negative correlations of this proportion, several exploratory analyses were undertaken. Scatter plots were drawn to show the relationship be-

tween teacher attitude scores and mean student scores on each of the three parts of the problem-solving test. The scatter plots and accompanying regression lines did, indeed, verify the negative nature of these relationships. Since these correlations were based on a relatively small and nonrandom sample, and since the attitudes of all teachers were favorable and the variance in scores was slight, the negative relationships were judged an artifact of this particular population.

Findings for Part II

The second part of the study was directed at questions 6 and 7 posed earlier in this chapter. The basic plan for Part II involved problem-solving testing at two different times with an intervening "treatment" period. Only the DMP sample of teachers and students was involved in this part of the study.

Findings for Question 6

The sixth question of the study was: Do fourth-grade teachers' attitudes toward problem-solving affect their students' problem-solving performance, or is the effect of the opposite nature? The cross-lagged panel correlational technique recommended by Campbell and Stanley (1963) was used for this part of the study, since simple correlational procedures cannot answer questions of cause and effect. As shown in Table 4 the correlations between student problem-solving performance at Time 1 and teacher problem-solving attitude at Time 2 were significantly more positive than the correlations between teacher attitude at Time 1 and student performance at Time 2. Thus, initial mean student problem-solving performance seemed to have a greater effect on final teacher attitudes than initial teacher attitudes had on final mean student problem-solving performance.

Cross-lagged panel correlations were also calculated for the data grouped by sex of students. The same directional relationships as in the total

Table 4

Cross-lagged Correlations: Time 1 Teacher Attitude with Time 2 Student Performance (r_{12}) and Time 2 Teacher Attitude with Time 1 Student Performance (r_{21})

Cross-lagged correlations	Comprehension	Application	Problem solving
r_{12}	-.72	-.72	-.69
r_{21}	-.25*	-.50*	-.53*

*Significant at $p < .01$ as determined by Fisher Z-transformation (see Hays, 1973).

sample were noted for girls, but for boys, the relationship was apparent only for the comprehension and application parts of the problem-solving test.

Findings for Question 7

The final question of the study was the following: Do fourth-grade teachers' attitudes toward problem solving affect their students' attitudes toward problem solving or is the effect of the opposite nature? The cross-lagged panel correlational technique was also employed to answer this question. Results are shown in Table 5. Correlations between teacher problem-solving attitude at Time 1 and student problem-solving attitude at Time 2 were significantly more positive than the correlations between teacher attitude at Time 2 and student attitude at Time 1. Thus, initial teacher attitude seemed to have a greater effect on final student attitude than initial student attitude had on final teacher attitude.

When cross-lagged correlations were calculated on the data grouped by sex of student, the same directional relationships held between teacher attitude and student attitude for boys and girls separately as held for the total sample.

Table 5

Cross-lagged Correlations: Time 1 Teacher Attitude with Time 2 Student Attitude (r_{12}) and Time 2 Teacher Attitude with Time 1 Student Attitude (r_{21})

Cross-lagged correlations	Attitude I	Attitude II	Total
r_{12}	.29	-.03	.13
r_{21}	-.47*	-.30*	-.37*

*Significant at $p < .01$ as determined by Fisher Z-transformation (see Hays, 1973).

Implications, Limitations, and Recommendations

Information-oriented research, such as the present study, provides insight into specific relationships between curriculum variables and suggests directions for additional studies. This section of the chapter, then, presents the implications and limitations of the study along with recommendations for future research.

Student Problem-solving Attitudes

If students in this study are reflective of those in a larger population, then most fourth-grade students do, indeed, possess favorable attitudes toward problem solving. Though not a random sample, the relatively large number of participating students strengthens the generalizability of the findings.

The problem-solving attitude scale developed for the study needs further validation with other elementary school populations. An interesting follow-up to the present study would be an observational investigation to determine if students actually possess the kinds of problem-solving behaviors which their responses to the problem-solving attitude scale indicate.

Teacher Problem-solving Attitudes

All teachers in the sample for the study indicated favorable attitudes toward problem solving, but, because there were only 30 of them, this finding may not be indicative of the larger population. Therefore, the teacher problem-solving attitude scale needs more extensive validation with other populations. The scale also could be used with prospective elementary school teachers to determine their attitudes towards mathematical problem solving.

Student Problem-solving Performance

The findings of the study indicate that fourth-grade students perform reasonably well on the first two parts of a test of mathematical problem solving which provides measures of comprehension, application, and problem solving. However, most students did not perform well on the third part of the test, a measure of problem-solving performance. The test by Romberg and Wearne (Wearne, 1976) holds promise as a viable tool for providing information to teachers and other school personnel about the problem-solving capabilities of students. This test can help diagnose student difficulties in comprehension, application, and problem solving. Once problem areas are diagnosed, teachers can plan remedies.

The fact that there were no significant differences between the problem-solving performance of boys and girls in this study indicates that teachers need not vary teaching techniques for the sexes. However, the fact that DMP students performed significantly better than non-DMP students on the comprehension and application portions of the test suggests that factors within the DMP program produce this effect. It would be interesting to determine whether similar differences exist in other populations of DMP and non-DMP students.

Student Problem-solving Attitudes and Performance

The significant and stable positive relationships found between student problem-solving attitude and student problem-solving performance suggest that the relationships between attitude and performance are the same for problem solving as for mathematics in general. Because of these positive relationships, it seems wise to foster favorable student reactions and sentiments toward all aspects of mathematical problem solving.

Teacher Problem-solving Attitude and Student Problem-solving Performance

The somewhat inconsistent findings in the relationships between teacher problem-solving attitude and student problem-solving performance,

when coupled with the relatively small sample of classes upon which the findings were based, suggest the need for gathering similar data from other similar populations. This suggestion also is based upon the surprising negative correlations that appeared in the DMP sample of the study. Clearly, more evidence is needed before definitive judgments can be made.

Cause and Effect Relationships Between Teacher Attitude and Student Attitude and Performance

Though calls for replication of research studies are easily made, the findings of the second part of the study obviously demand such efforts. If the directional relationship is one way for teacher attitude and student performance, and the opposite direction for teacher attitude and student attitude, teachers should be aware of this situation. If this directional influence is peculiar to a particular population, then knowledge of that fact would be beneficial.

The cross-lagged panel correlational technique (Campbell & Stanley, 1963) holds promise as a valuable research design for inferring the cause and effect relationships between such variables as attitude and performance. As a follow-up to the present investigation, the author suggests that an improved use for the cross-lagged technique might involve initial problem-solving testing with students and teachers near the start of the school year and again at mid-year; this plan would reduce the confounding teacher-pupil effect occurring when initial testing is done several weeks into the school year.

Concluding Remarks

The study reported in this chapter investigated selected noncognitive factors and the mathematical problem-solving performance of fourth-grade children. As is often the case, the results have raised more questions than they have answered. In the author's opinion, the most important findings of the study are: (a) fourth-grade students and teachers seem to possess favorable attitudes toward mathematical problem solving; (b) fourth-grade students perform satisfactorily on comprehension and application items, but not on the problem-solving items of a three-part mathematical problem-solving test; and (c) there seems to be a significant and stable positive relationship between student mathematical problem-solving performance and student problem-solving attitude. The other findings of the study are important, but must be viewed as tentative until validated with additional research.

References

Ahmann, J. S. *Consumer math: Selected results from the first national assessment of mathematics* (N.AEP Mathematics Report No. 04-MA-02). Washington, D.C.: U.S. Government Printing Office, 1975.

Aiken, L. C. Personality correlates of attitude toward mathematics. *Journal of Educational Research*, 1963, *56*, 476-480.

Aiken, L. R. Attitudes toward mathematics. In J. W. Wilson & L. R. Carry (Eds.), *Studies in mathematics* (Vol. XIX). Stanford, CA: School Mathematics Study Group, 1969.

Aiken, L. R. Affective factors in mathematics learning: Comments on a paper by Neale and a plan for research. *Journal for Research in Mathematics Education*, 1970, *1*, 251-255.

Aiken, L. R. Biodata correlates of attitudes toward mathematics in three age and two sex groups. *School Science and Mathematics*, 1972, *72*, 386-395.

Aiken, L. R., & Dreger, R. M. The effect of attitudes on performance in mathematics. *Journal of Educational Psychology*, 1961, *52*, 19-24.

Alexander, V. E. Seventh graders' ability to solve problems. *School Science and Mathematics*, 1960, *60*, 603-606.

Allport, G. W. Attitudes. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: John Wiley & Sons, 1967.

American Psychological Association. *Standards for educational and psychological tests and manuals*. Washington, D.C.: Author, 1966.

Anderson, R. C. Learning in discussions: A resumé of the authoritarian-democratic studies. In W. W. Charters & N. L. Gage (Eds.), *Readings in the social psychology of education*. Boston: Allyn & Bacon, Inc., 1963.

Anglin, L., Meyer, R., & Wheeler, J. *The relationship between spatial ability and mathematics achievement at the fourth- and sixth-grade levels* (Working Paper 115). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1975.

Anttonen, R. G. A longitudinal study in mathematics attitude. *Journal of Educational Research*, 1969, *62*, 467-471.

Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory* (2nd ed.). New York: Harper & Row, 1960.

Association for Supervision and Curriculum Development. *Humanizing education: The person in the process.* Washington, D.C.: Author, National Education Association, 1967.

Ausubel, D. P. *The psychology of meaningful verbal learning.* New York: Grune & Stratton, 1963.

Ausubel, D. P. *Educational psychology: A cognitive view.* New York: Holt, Rinehart, & Winston, 1968.

Baker, F. B. Generalized item and test analysis program (GITAP). *Educational and Psychological Measurement,* 1963, *23,* 187-190.

Baker, F. B. *FORTAP: A FORTRAN test analysis package.* Department of Educational Psychology, University of Wisconsin, 1969.

Bales, R. F., & Borgatta, E. F. Size of group as a factor in the interaction profile. In A. P. Hare, E. F. Borgatta, & R. F. Bales (Eds.), *Small groups, studies in social interaction.* New York: Alfred A. Knopf, 1961.

Balow, I. H. Reading and computation ability as determinants of problem solving. *Arithmetic Teacher,* 1964, *11,* 18-22.

Barrett, E. S. An analysis of verbal reports of solving spatial problems as an aid in defining spatial factors. *Journal of Psychology,* 1953, *36,* 17-25.

Beck, A., Bleicher, M. N., Crowe, D. W., & Libeskind, S. Teacher's manual. In A. Beck, M. N. Bleicher, & D. W. Crowe (Eds.), *Excursions into mathematics.* New York: Worth Publishers, Inc., 1970.

Begle, E. G., & Wilson, J. W. Evaluation of mathematics programs. In E. G. Begle (Ed.), *Mathematics education.* Chicago: National Society for the Study of Education, 1970.

Beldin, H. O. *A study of selected arithmetic verbal problem solving skills among high and low achieving sixth grade children.* Unpublished doctoral dissertation, Syracuse University, 1960.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. *Differential aptitude tests, form T* (4th ed.). New York: The Psychological Corporation, 1972.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. *Differential aptitude tests: Administrator's handbook.* New York: The Psychological Corporation, 1973.

Berry, P. C. *An exploration of the interrelations among nonintellectual predictors of achievement in problem solving* (Technical Report 4). New Haven, CT: Department of Psychology and Department of Industrial Administration, Yale University, 1958.

Berry, P. C. *A second exploration of the interrelations among nonintellectual predictors of achievement in problem solving* (Technical Report 5).

New Haven, CT: Department of Psychology and Department of Industrial Administration, Yale University, 1959.

Blake, R. N. The effect of problem context upon the problem solving processes used by field dependent and independent subjects: A clinical study (Doctoral dissertation, University of British Columbia, 1976). *Dissertation Abstracts International*, 1977, *37*, 4191-4192A.

Block, J. H. (Ed.). *Mastery learning theory and practice*. New York: Holt, Rinehart & Winston, Inc., 1971.

Bock, R. D., & Kolakowski, D. Further evidence of sex-linked major-gene influence on human spatial visualizing ability. *American Journal of Human Genetics*, 1973, *25*(1), 1-14.

Boe, B. L. A study of the ability of secondary pupils to perceive plane sections of selected solid figures. *Mathematics Teacher*, 1968, *61*, 415-421.

Bogolyubev, A. N. Work with words in the solution of arithmetic problems in elementary school. In J. Kilpatrick & I. Wirszup (Eds.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. VI). Chicago: University of Chicago, 1972.

Botsmanova, M. E. [The forms of pictorial visual aids in instruction in arithmetic problem solving.] In J. Kilpatrick & I. Wirszup (Eds.) & J. W. Teller (trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. VI). Stanford, CA: School Mathematics Study Group, Stanford University & Survey of Recent East European Mathematical Literature, University of Chicago, 1972. (Originally published, 1960.)

Bourne, L. E., & Battig, W. F. Complex processes. In J. Sidowski (Ed.), *Experimental methods and instrumentation in psychology*. New York: McGraw-Hill, 1966.

Brechting, M. C., & Hirsch, C. R. *The effects of small group-discovery learning on student achievement and attitudes in calculus*. Unpublished manuscript, 1977.

Broder, L. J., & Bloom, B. S. *Problem solving processes of college students* (Supplementary Educational Monographs, No. 73). Chicago: University of Chicago, 1950.

Brown, E. D. Arithmetical understandings and attitudes toward arithmetic of experienced and inexperienced teachers (Doctoral dissertation, University of Nebraska, 1961. *Dissertation Abstracts International*, 1962, *22*, 775.

Brown, S. I. Learning by discovery in mathematics: Rationale, implementation, and misconceptions. *Educational Theory*, Summer 1971, *21*, 232-260.

Brownell, W. A. Problem solving. *The psychology of learning* (Yearbook 41, Part II, National Society for the Study of Education). Chicago: University of Chicago, 1942.

Bruner, J. S. Some theorems on instruction illustrated with reference to mathematics. In E. R. Hilgard (Ed.), *Theories of learning and instruction* (Yearbook 63, Part I, National Society for the Study of Education). Chicago: University of Chicago, 1964.

Buchoff, D. E. *A comparative study of random and homogeneous grouping of paired individuals using programmed instructional materials in plane geometry*. Unpublished master's thesis, Univesity of Maryland, 1970.

Buck, R. C. A look at mathematics competitions. *American Mathematical Monthly*, 1959, *66*, 201-212.

Buck, R. C. Teaching machines and mathematics programs: Statement by R. C. Buck, *American Mathematical Monthly*, 1962, *69*(6), 561-564.

Buck, R. C. Goals for mathematics instruction. *American Mathematical Monthly*, 1965, *72*, 949-956.

Butcher, H. J. *Human intelligence*. London: Methuen & Co. Ltd., 1968.

Caezza, J. F. A study of teacher experience, knowledge of and attitude toward mathematics and the relationship of these variables to elementary school pupils' attitudes toward and achievement in mathematics (Doctoral dissertation, Syracuse University, 1969). *Dissertation Abstracts International*, 1970, *31*, 921A-922A.

Callahan, W. J. Adolescent attitudes toward mathematics. *Mathematics Teacher*, 1971, *64*, 751-755.

Cambridge Conference on School Mathematics. *Goals for school mathematics*. New York: Houghton Mifflin Company, 1963.

Campbell, D. F. Factorial comparison of arithmetic performance of boys in sixth and seventh grade. *Educational Research Monographs*, 1956, *20*(2), 1-25.

Campbell, D. T. The indirect assessment of social attitudes. *Psychological Bulletin*, 1950, *47*, 15-38.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Co., 1963.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Co., 1969.

Carey, G. L. *Reduction of sex differences in problem solving by improvement of attitude through group discussion* (Technical Report 9). Stanford, CA: Stanford University Department of Psychology, 1955.

Carey, G. L. Sex differences in problem solving performance as a function of attitude differences. *Journal of Abnormal and Social Psychology*, 1958, *56*, 256-260.

Carnegie Commission on Higher Education. *Opportunities for women in higher education*. New York: McGraw-Hill, 1973.

Carry, L. R. Patterns of mathematics achievement in grades 7 and 8: X-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 11). Stanford, CA: School Mathematics Study Group, Stanford University, 1970.

Carry, L. R., & Weaver, J. F. Patterns of mathematics achievement in grades 4, 5 and 6: X-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 10). Stanford, CA: School Mathematics Study Group, Stanford University, 1969.

Cartwright, D., & Zander, A. (Eds.). *Group dynamics: Research and theory* (3rd ed.). New York: Harper & Row, 1968.

Chakerian, G. D., Crabill, C. D., & Stein, S. K. *Geometry: A guided inquiry*. Boston: Houghton Mifflin Co., 1972.

Chase, C. I. The position of certain variables in the prediction of problem solving in arithmetic. *Journal of Educational Research*, 1960, *54*, 9-14.

Clarkson, S. P. [Problem solving processes of mentally retarded children.] In J. Kilpatrick, I. Wirszup, E. G. Begle, & J. W. Wilson (Eds. and trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. IX). Chicago: University of Chicago, 1975. (a)

Clarkson, S. P. Teaching mathematics to mentally retarded children. In J. Kilpatrick, I. Wirszup, E. G. Begle, & J. W. Wilson (Eds. and trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. X). Chicago: University of Chicago, 1975. (b)

Cleveland, G. A. A study of certain psychological and sociological characteristics as related to arithmetic achievement (Doctoral dissertation, Syracuse University, 1961). *Dissertation Abstracts International*, 1962, *22*, 2681-2682.

Cochran, W. G. Analysis of covariance: Its nature and uses. *Biometrics*, 1957, *13*, 261-281.

College Entrance Examination Board. *1972-73 advanced placement mathematics*. Princeton, NJ: Author, 1972.

College Entrance Examination Board Commission on Mathematics. *Program for college preparatory mathematics* (Report). New York: College Entrance Examination Board, 1959.

Cook, S. W., & Selltiz, C. A multiple-indicator approach to attitude measurement. *Psychological Bulletin*, 1964, *62*, 36-55.

Coon, L. H. Attitude: A rating scale for calculus. *Journal of Educational Research*, 1969, *63*, 173-177.

Cooper, J. B., & Pollack, D. The identification of prejudicial attitudes by the galvanic skin response. *Journal of Social Psychology*, 1959, *50*, 241-245.

Cooperative Test Division. *Sequential tests of educational progress*. Princeton, NJ: Educational Testing Service, 1956-1972.

Coopersmith, S. A method for determining types of self-esteem. *Journal of Abnormal and Social Psychology*, 1959, *59*, 87-94.

Corcoran, M., & Gibb, G. Appraising attitudes in the learning of mathematics. In *Evaluation in mathematics* (Yearbook 26). Washington, D.C.: The National Council of Teachers of Mathematics, 1961.

Covington, M. V. A childhood attitude inventory for problem solving. *Journal of Educational Measurement*, 1966, *3*, 234.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

Crosswhite, F. J. Correlates of attitudes toward mathematics. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 20). Stanford, CA: School of Mathematics Study Group, Stanford University, 1972.

Cummins, K. B. A student experience-discovery approach to the teaching of calculus (Doctoral dissertation, Ohio State University, 1958). *Dissertation Abstracts International*, 1959, *19*, 2292. (University Microfilms No. 59-367)

Cummins, K. B. A student experience-discovery approach to the teaching of calculus. *Mathematics Teacher*, 1960, *53*(3), 162-170.

Crueton, E. E. Reliability and validity. Basic assumptions and experimental designs. *Educational and Psychological Measurement*, 1965, *25*, 327-346.

Dalton, R. M. Thinking patterns in solving certain word problems by ninth grade general mathematics students: An exploratory study in problem solving (Doctoral dissertation, University of Tennessee, 1974). *Dissertation Abstracts International*, 1975, *35*, 5526B. (University Microfilms No. 75-11,155)

Dancis, J. *Linear algebra.* Unpublished manuscript.

Davidson, D. Learning mathematics in a group situation. *Mathematics Teacher,* 1974, *67*(2), 101-106.

Davidson, N. A. *The small group-discovery method of mathematics instruction as applied in calculus* (Technical Report 168). Madison: Wisconsin Research and Development Center for Cognitive Learning, October 1971. (a)

Davidson, N. A. The small group-discovery method as applied in calculus instruction. *American Mathematical Monthly,* August-September 1971, 789-791. (b)

Davidson, N. A. Motivation of students in small group learning of mathematics. *Frostburg State College Journal of Mathematics Education,* 1976, *11,* 1-18.

Davidson, N. A., Agreen, L., & Davis, C. Small group learning in junior high school mathematics. *School Science and Mathematics,* January 1978, 23-30.

Davidson, N. A., & Gulick, F. *Abstract algebra: An active learning approach.* Boston: Houghton Mifflin Co., 1976.

Davidson, N. A., & Leach, R. *Calculus: A student discovery approach.* Book in preparation, 1973.

Davidson, N. A., McKeen, R., & Eisenberg, T. Curriculum construction with student input. *Mathematics Teacher,* March 1973, 271-275.

Davidson, N. A., & Urion, D. Student achievement in small group instruction versus teacher-centered instruction in mathematics. To appear in the *Two-year College Mathematics Journal.*

Deutsch, M. The effects of cooperation and competition upon group process. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory* (2nd ed.). New York: Harper & Row, 1960.

Deutsch, M., & Gerard, H. B. A study of normative and informational social influences upon individual judgment. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory* (2nd ed.). New York: Harper & Row, 1960.

Dewey, J. *Experience and education.* New York: Collier Books Paperback Edition, 1963. (Originally published, 1938.)

Dewey, J. *Democracy and education.* New York: Free Press Paperback Edition, 1966. (Originally published, 1916.)

Dodson, J. W. Characteristics of successful problem solvers (Doctoral dissertation, University of Georgia, 1970). *Dissertation Abstracts International*, 1971, *31*, 5928A. (University Microfilms No. 71-13,048)

Dodson, J. W. Characteristics of successful insightful problem solvers. In J. W. Wilson & E. G. Begle (Eds.), *NLSMA Reports* (No. 31). Stanford, CA: School Mathematics Study Group, Stanford University, 1972.

Donohue, J. C. Factorial comparison of arithmetic problem solving ability of boys and girls in seventh grade. *Educational Research in Monographs*, 1957, *20*(2), 1-30.

Downie, N. M., & Heath, R. W. *Basic statistical methods* (3rd ed.). New York: Harper & Row, 1970.

Draper, N. R., & Smith, H. *Applied regression analysis*. New York: John Wiley & Sons, 1966.

Droege, R. C. Sex differences in aptitude maturation during high school. *Journal of Counseling Psychology*, 1967, *14*, 407-411.

Duncker, K. On problem solving. *Psychological Monographs*, 1945, *58*, 1-113. (Whole No. 270)

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. *Metropolitan achievement tests: Advanced form F*. New York: Harcourt, Brace, Jovanovich, 1970.

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. *Metropolitan achievement tests: Advanced form F*. New York: Harcourt, Brace, Jovanovich, 1971. (a)

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. *Metropolitan achievement tests: Teacher's handbook*. New York: Harcourt, Brace, Jovanovich, 1971. (b)

Dutton, W. H. Attitude change of prospective elementary school teachers toward arithmetic. *Arithmetic Teacher*, 1962, *9*, 418-424.

Educational Services, Incorporated. *Goals for school mathematics* (Report of the Cambridge Conference on School Mathematics). Boston: Houghton Mifflin Co., 1963.

Edwards, R. M. Factorial comparison of arithmetic performance of girls and boys in the sixth grade. *Educational Research Monographs*, 1957, *20*(7), 1-38.

Egan, D. E., & Green, J. G. Acquiring cognitive structure by discovery and rule learning. *Journal of Educational Psychology*, 1973, *64*, 85-97.

Eisenberg, T. A. *The integration of modified learner-generated sequences into the development of a behaviorally stated learning hierarchy, as applied*

in mathematics curricula construction. Unpublished doctoral dissertation, University of Maryland, 1970.

Emm, M. E. A factorial study of problem solving ability of fifth grade boys. *Educational Research Monographs*, 1959, *22*(1), 1-51.

Engelhard, M. D. *An experimental study of arithmetic problem solving ability of fourth grade girls*. Unpublished doctoral dissertation, The Catholic University of America, 1955.

Ernst, G. W., & Newell, A. *GPS: A case study in generalizability and problem solving*. New York: Academic Press, 1969.

Evans, R. F. A study of the reliabilities of four arithmetic scales and an investigation of component mathematics attitudes (Doctoral dissertation, Case Western Reserve University, 1971). *Dissertation Abstracts International*, 1972, *32*, 3086A-3087A.

Evanson, J. Madison Academic Computing Center, Madison, WI: Personal communication, July 1975.

Faust, C. E. A study of the relationship between attitude and achievement in selected elementary school subjects (Doctoral dissertation, State University of Iowa, 1963). *Dissertation Abstracts International*, 1963, *23*, 2752-2753.

Fedon, J. P. The role of attitude in learning arithmetic. *Arithmetic Teacher*, 1958, *5*, 304-310.

Fehr, H. F., Fey, J. T., & Hill, T. J. *Unified mathematics course II*. Menlo Park, CA: Addison-Wesley, 1972. (a)

Fehr, H. F., Fey, J. T., & Hill, T. J. *Unified mathematics course III*. Menlo Park, CA: Addison-Wesley, 1972. (b)

Fennema, E. H. *Mathematics, spatial ability and the sexes*. Paper presented at the meeting of the American Educational Research Association, Chicago, 1974.

Fennema, E. H., & Sherman, J. A. *Sex-related differences in mathematics achievement and related factors: A further study*. Manuscript submitted for publication, 1976.

Fennema, E. H., & Sherman, J. A. Sex-related differences in mathematics spatial visualization and affective factors. *American Educational Research Journal*, 1977, *14*, 51-71.

Fey, J. Classroom teaching of mathematics. *Review of Educational Research*, 1969, *39*, 535-551.

Finkbeiner, D. T., Neff, J. D., & Williams, S. I. 1969 advanced placement examination in mathematics: Complete and unexpurgated. *Mathematics Teacher*, 1971, *64*, 497-516.

Flaherty, E. G. Cognitive processes used in solving mathematical problems (Doctoral dissertation, Boston University, 1973). *Dissertation Abstracts International*, 1973, *34*, 1767A. (University Microfilms No. 73-23,562)

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycroft, M. F., Orr, D. B., Goldberg, I., & Neyman, C. A. *The American high school student today*. Pittsburgh: University of Pittsburgh, 1964.

Foster, T. E. The effect of computer programming experiences on student problem solving behaviors in eighth-grade mathematics (Doctoral dissertation, University of Wisconsin-Madison, 1972). *Dissertation Abstracts International*, 1973, *33*, 4239A. (University Microfilms No. 72-31,527)

French, J. W. The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement*, 1965, *25*, 9-28.

French, J. W., Ekstrom, R. B., & Price, L. A. *Kit of reference tests for cognitive factors* (Rev. 1963). Princeton, NJ: Educational Testing Service, 1969. (a)

French, J. W., Ekstrom, R. B., & Price, L. A. *Manual for kit of reference tests for cognitive factors*. Princeton, NJ: Educational Testing Service, 1969. (b)

Fruchter, B. Measurement of spatial abilities: History and background. *Educational and Psychological Measurement*, 1954, *14*, 387-395.

Fuller, S. F. *A study of problem solving methods of students with and without the constraint of a time limit*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Atlanta, April 1972.

Gagné, R. M. Educational objectives and human performance. In J. D. Krumboltz (Ed.), *Learning and the educational process*. Chicago: Rand McNally & Co., 1965.

Gallicchio, A. *The effects of brainstorming in small group mathematics classes*. Unpublished doctoral dissertation, University of Maryland, 1976.

Gallo, D. P. Problem solving: A test of two aspects (Doctoral dissertation, State University of New York at Stony Brook, 1974). *Dissertation Abstracts International*, 1975, *35*, 4649-4650B. (University Microfilms No. 75-5373)

Garai, J. E., & Scheinfeld, A. Sex differences in mental and behavioral traits. *Genetic Psychology Monographs*, 1968, *77*, 169-299.

Gee, B. C. Attitudes toward mathematics and basic mathematical understanding of prospective elementary school teachers at Brigham Young University (Doctoral dissertation, Oregon State University, 1966). *Dissertation Abstracts International*, 1966, *26*, 6528.

Gimmestad, B. J. An exploratory study of the processes used by community college students in mathematical problem solving (Doctoral dissertation, University of Colorado at Boulder, 1976). *Dissertation Abstracts International*, 1977, *37*, 7590A. (University Microfilms No. 77-11,300)

Goldberg, D. J. The effects of training in heuristics on the ability to write proofs in number theory (Doctoral dissertation, Columbia University, 1974). *Dissertation Abstracts International*, 1975, *35*, 4989B. (University Microfilms No. 75-7,836)

Goodman, L. A. On partitioning X^2 and detecting partial association in three-way contingency tables. *Journal of the Royal Statistical Society*, 1969, *31*(Series B), 486-498.

Goodman, L. A. The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 1970, *65*, 226-256

Goodman, L. A. Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association*, 1971, *66*, 339-344.

Grant, S. *The effects of three kinds of group formation using FIRO-B compatibility, sociometric choice with group dynamics exercises, and in-class choice on mathematics classes taught by the small group discovery method.* Unpublished doctoral dissertation, University of Maryland, 1975.

Guilford, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

Guttman, L. A new approach to factor analysis: The radex. In P. F. Lazarfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, IL: The Free Press, 1954.

Hadamard, J. *The psychology of invention in the mathematical field.* New York: Dover, 1954.

Hall, T. R. A study of situational problem solving by gifted high school mathematics students (Doctoral dissertation, Georgia State University, 1976). *Dissertation Abstracts International*, 1976, *37*, 906-907A. (University Microfilms No. 76-16,964)

Hallgren, S. O. Effects of verbalization, tempo, and problem size on problem solving (Doctoral dissertation, Michigan State University, 1976). *Dissertation Abstracts International*, 1976, *37*, 3111B. (University Microfilms No. 76-27,103)

Handler, J. R. An exploratory study of the spatial visualization abilities and problem solving processes exhibited by high school mathematics students while solving a set of geometric problems (Doctoral dissertation, University of Tennessee, 1976). *Dissertation Abstracts International*, 1977, *37*, 7008A. (University Microfilms No. 77-10,770)

Harris, C. W. Some Rao-Guttman relationships. *Psychometrika*, 1962, *27*, 247-263.

Harris, C. W., & Kaiser, H. F. Oblique factor analytic solution by orthogonal transformations. *Psychometrika*, 1964, *29*, 347-362.

Harris, M. L. *Some methodological suggestions for construction of an objective measurement instrument* (Technical Memo M-1968-2). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1968.

Harris, M. L., & Harris, C. W. A factor analytic interpretation strategy. *Educational and Psychological Measurement*, 1971, *31*, 589-606.

Harris, M. L., & Harris, C. W. A structure of concept attainment abilities. *Wisconsin Monograph Series*. Madison: Wisconsin Research and Development Center for Cognitive Learning, 1973.

Hatfield, L. L. *Heuristic emphasis in the instruction of mathematical problem solving: Rationale and research*. Paper presented at Research Workshop on Mathematical Problem Solving, Georgia Center for the Study of Learning and Teaching Mathematics (GCSLTM), University of Georgia, Athens, May 1975.

Hatfield, L. L. The problem solving project of the Georgia Center for the Study of Learning and Teaching Mathematics. Paper presented at 3rd International Congress on Mathematical Education, Karlsruhe, Germany, August 1976.

Hays, W. L. *Statistics*. New York: Holt, Rinehart & Winston, 1963.

Hays, W. L. *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston, Inc., 1973.

Heimer, R. T., & Lottes, J. J. Toward a theory of sequencing: An integrated program of research. *Journal for Research in Mathematics Education*, 1973, *4*(2), 85-93.

Hein, P. *Grooks*. New York: Doubleday, 1966.

Hieronymus, A. N., & Lindquist, E. F. *Iowa test of basic skills, form 6.* Boston: Houghton Mifflin Co., 1971.

Higgins, J. L. A new look at heuristic teaching. *Mathematics Teacher,* 1971, *64,* 487-495.

Hilbert, D. Mathematical problems. In *Archives of mathematics and physics,* 1906.

Hildebrand, R. *Measuring student attitudes toward small group instruction in mathematics.* Paper presented at the Washington meeting of the National Council of Teachers of Mathematics, Washington, D.C., March 23, 1975.

Hilton, T. L., & Berglünd, G. W. Sex differences in mathematics achievement — a longitudinal study. *Journal of Educational Research,* 1974, *67,* 231-237.

Hobson, J. R. Sex differences in primary mental abilities. *Journal of Educational Research,* 1947, *41,* 126-132.

Hoffman, L. R., & Maier, N. R. F. Social factors influencing problem solving in women. *Journal of Personality and Social Psychology,* 1966, *4,* 382-390.

Hollowell, K. A. A flow chart model of cognitive processes in mathematical problem solving (Doctoral dissertation, Boston University, 1977). *Dissertation Abstracts International,* 1977, *37,* 7666-7667A. (University Microfilms No. 77-11,363)

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika,* 1941, *12,* 153-160.

Hubert, L. J. Monotone invariant clustering procedures. *Psychometrika,* 1973, *38,* 49-55.

Hughes, B. B. Heuristic teaching in mathematics. *Educational Studies in Mathematics,* 1974, *5,* 291-299.

The Instructional Objectives Exchange. *Instructional objectives exchange — Mathematics K-6.* Los Angeles: Author, 1970.

Jarvis, O. T. Boy-girl differences in elementary school arithmetic. *School Science and Mathematics,* 1964, *64,* 657-659.

Johnson, H. C. Problem-solving in arithmetic: A review of the literature. *Elementary School Journal,* 1944, *44,* 396-403, 476-482.

Johnson, J. T. On the nature of problem solving in arithmetic. *Journal of Educational Research,* 1949, *43,* 110-115.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 241-254.

Jordy, G. *Small group-discovery lessons for SSMCIS II and III with an exploratory school-based study of their use.* Unpublished doctoral dissertation, University of Maryland, 1976.

Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 1967, *32*, 443-482.

Kabanova-Meller, E. N. [The role of the diagram in the application of geometric theorems.] In J. Kilpatrick & I. Wirszup (Eds.) & M. Ackerman (trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. IV). Stanford, CA: School Mathematics Study Group, Stanford University and Survey of Recent East European Mathematical Literature, University of Chicago, 1970. (Originally published, 1950.)

Kagan, J., & Kagan, N. Individual variation in cognitive processes. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (3rd ed., Vol. 1). New York: John Wiley & Sons, 1970.

Kagan, J., & Moss, H. A. *Birth to maturity.* New York: John Wiley & Sons, 1962.

Kaiser, H. F. The varimax criterion for analytic relation in factor analysis. *Psychometrika*, 1958, *23*, 187-200.

Kaiser, H. F., & Caffrey, J. Alpha factor analysis. *Psychometrika*, 1965, *30*, 1-14.

Kantowski, M. G. *Processes involved in mathematical problem solving.* Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Denver, 1974.

Kantowski, M. G. Analysis and synthesis of problem solving methods. In J. Kilpatrick, I. Wirszup, E. G. Begle, & J. W. Wilson (Eds. and trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. XI). Chicago: University of Chicago, 1975. (a)

Kantowski, M. G. *The teaching experiment and Soviet studies of problem solving.* Paper presented at the Research Workshop on Mathematical Problem Solving, Georgia Center for the Study of Learning and Teaching Mathematics (GCSLTM), University of Georgia, Athens, May 1975. (b)

Kantowski, M. G. Processes involved in mathematical problem solving. *Journal for Research in Mathematics Education*, 1977, *8*, 163-180.

Katz, D. The functional approach to the study of attitudes. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: John Wiley & Sons, 1967.

Kelley, T., Madden, R., Gardner, E., & Rudman, H. *Directions for administering intermediate I battery* . New York: Harcourt, Brace & World, Inc., 1964.

Kendall, M. G. *Rank correlation methods*. London: Charles Griffin & Co., Ltd., 1955.

Kenney, P. *Effects of group cooperation stimulated by competition between groups as a motivating technique in a ninth grade mathematics classroom*. Unpublished doctoral dissertation, Catholic University of America, 1974.

Kilpatrick, J. Analyzing the solution of word problems in mathematics: An exploratory study (Doctoral dissertation, Stanford University, 1967). *Dissertation Abstracts International*, 1968, *28*, 4380A. (University Microfilms No. 68-6,442)

Kilpatrick, J. Problem solving and creative behavior in mathematics. In J. W. Wilson & L. R. Carry (Eds.), *Studies in mathematics* (Vol. XIX). Stanford, CA: School Mathematics Study Group, Stanford University, 1969.

Kilpatrick, J. Problem solving in mathematics. *Review of Educational Research*, 1970, *39*, 523-534.

Kilpatrick, J. *Variables and methodologies in research on problem solving*. Paper presented at the Research Workshop on Mathematical Problem Solving, Georgia Center for the Study of Learning and Teaching Mathematics (GCSLTM), University of Georgia, Athens, Georgia, May 1975.

Kilpatrick, J., & McLeod, G. K. Patterns of mathematics achievement in grade 9: Y-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 13). Stanford, CA: School Mathematics Study Group, Stanford University, 1971. (a)

Kilpatrick, J., & McLeod, G. K. Patterns of mathematics achievement in grade 11: Y-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 15). Stanford, CA: School Mathematics Study Group, Stanford University, 1971. (b)

Kilpatrick, J . & Wirszup, I. (Eds. and trans.). The structure of mathematical abilities. *Soviet studies in the psychology of learning and teaching mathematics* (Vol. II). Stanford, CA: School Mathematics Study Group, Stanford University, 1969. (a)

Kilpatrick, J., & Wirszup, I. (Eds. and trans.). Problem solving in algebra and arithmetic. *Soviet studies in the psychology of learning and teaching mathematics* (Vol. III). Stanford, CA: School Mathematics Study Group, Stanford University, 1969. (b)

Kilpatrick, J., & Wirszup, I. (Eds. and trans.). Problem solving in geometry. *Soviet studies in the psychology of learning and teaching mathematics* (Vol. IV). Stanford, CA: School Mathematics Study Group, Stanford University, 1970.

Kilpatrick, J., & Wirszup, I. (Eds. and trans.) Instruction in problem solving. *Soviet studies in the psychology of learning and teaching mathematics* (Vol. VI). Stanford, CA: School Mathematics Study Group, Stanford University, 1972.

King, I. L. *A formative development of a unit on proof for use in the elementary school* (Technical Report 111). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1970.

Kingsbury, D. *An experiment in education.* Unpublished manuscript, 1963. (Available from the Mathematics Department, McGill University, Montreal 2, P.Q., Canada.)

Kliebhan, M. C. *An experimental study of arithmetic problem solving ability of 6th grade boys.* Unpublished doctoral dissertation, Catholic University of America, 1955.

Kline, M. *Why Johnny can't add: The failure of the new math.* New York: Vintage Books, 1973.

Klingbeil, D. *An examination of the effects of group testing in mathematics courses taught by the small group-discovery method.* Unpublished doctoral dissertation, University of Maryland, 1974.

Klingbeil, D., & Davidson, N. *Student performance in a program of small group testing in mathematics.* Manuscript submitted for publication.

Knaup, J., Smith, L. J., Shoecraft, P., & Warkentin, G. D. *Patterns and systems of elementary mathematics.* Boston: Houghton Mifflin Co., 1977.

Kolb, D. A., Rubin, I. M., & McIntyre, J. M. *Organizational psychology: An experimental approach* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1974.

Krutetskii, V. A. [*The psychology of mathematical abilities in school children*] (J. Kilpatrick & I. Wirszup, Eds. and J. Teller, trans.). Chicago: University of Chicago, 1976. (Originally published, 1968.)

Krutetskii, V. A. [An investigation of mathematical abilities in school children.] In J. Kilpatrick & I. Wirszup (Eds.), *Soviet studies in the psy-*

chology of learning and teaching mathematics (Vol. II). Stanford, CA: School Mathematics Study Group, Stanford University, 1969.

Kulm, G. (Ed.). *Catalogue of problem solving instruments.* Athens, GA: Georgia Center for the Study of Learning and Teaching Mathematics, 1977.

Lankton, R. S. *Lankton first-year algebra test* (Rev. ed.). New York: Harcourt, Brace, Jovanovich, 1965.

Larsen, C. M. The heuristic standpoint in the teaching of elementary calculus (Doctoral dissertation, Stanford University, 1960). *Dissertation Abstracts International,* 1961, *21*(9), 2632-2633. (University Microfilms No. 60-6,740)

Leach, R., & Davidson, N. *Calculus via small group activities.* Unpublished manuscript, 1978.

Ledbetter, M. *An investigation of the interactions of student ability profiles and instruction in heuristic strategies with problem solving performance and problem sorting schemes.* Paper presented at the annual meeting of the National Council of Teachers of Mathematics, San Diego, CA, April 1978.

Lee, K. S. An exploratory study of fourth-graders' heuristic problem solving behavior (Doctoral dissertation, University of Georgia; 1977). *Dissertation Abstracts International,* 1977, *38,* 4004A. (University Microfilms No. 77-29,779)

Leggette, E. C. The effect of a structured problem solving process on the problem solving ability of capable but poorly prepared college freshmen in mathematics (Doctoral dissertation, Rutgers University, 1973). *Dissertation Abstracts International,* 1974, *34,* 3838A. (University Microfilms No. 73-32,223)

Leithold, L. *The calculus with analytic geometry* (1st ed.). New York: Harper & Row, 1968.

Lester, F. K. Developmental aspects of human problem solving in a simple mathematical system via computer assisted instruction (Doctoral dissertation, Ohio State University, 1972). *Dissertation Abstracts International,* 1973, *33,* 4178A. (University Microfilms No. 73-2,047)

Lester, F. K. *Developmental aspects of mathematical problem solving.* Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Atlantic City, April 1974.

Libeskind, S. A simple constructive proof for two identities. *Mathematics Teacher,* 1968, *61*(3), 259-263.

Libeskind, S. *A development of a unit on number theory for use in high school, based on a heuristic approach.* Unpublished doctoral dissertation, University of Wisconsin, 1971.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 1932, *140*, 44-53.

Lindgren, H. C., Silva, I., Faraco, I., & DaRocha, N. S. Attitudes toward problem solving as a function of success in arithmetic in Brazilian elementary schools. *Journal of Educational Research*, 1964, *58*, 44-45.

Lindquist, E. F., & Hieronymus, A. N. *Iowa tests of basic skills.* Boston: Houghton Mifflin Co., 1964.

Lindquist, E. F., & Hieronymus, A. N. *Iowa tests of basic skills.* Boston: Houghton Mifflin Co., 1973.

Lingoes, J. C. *The Guttman-Lingoes non-metric program series.* Ann Arbor, MI: Maththesis Press, 1973.

Linville, W. J. *The effects of syntax and vocabulary upon the difficulty of verbal arithmetic problems with fourth grade students.* Unpublished doctoral dissertation, Indiana University, 1969.

Lipson, S. H. The effects of teaching heuristics to student teachers in mathematics (Doctoral dissertation, Columbia University, 1972). *Dissertation Abstracts International*, 1972, *33*, 2221-2222A. (University Microfilms No. 72-30,334)

Loevinger, J. The attenuation paradox in test theory. *Psychological Bulletin*, 1954, *51*, 493-504.

Loomer, N. J. A multidimensional exploratory investigation of small group-heuristic and expository learning in calculus (Doctoral dissertation, University of Wisconsin, 1976). *Dissertation Abstracts International*, 1976, *37*, 2697A. (University Microfilms No. 76-18,892)

Lord, F. M., & Novick, M. R. *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Lord, F. M., & Novick, M. R. *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley, 1974.

Lorge, I., Thorndike, R. L., & Hagen, E. *Lorge-Thorndike intelligence tests.* Boston: Houghton Mifflin Co., 1966.

Lucas, J. F. An exploratory study on the diagnostic teaching of heuristic problem solving strategies in calculus (Doctoral dissertation, University of Wisconsin-Madison, 1972). *Dissertation Abstracts International*, 1972, *32*, 6825A. (University Microfilms No. 72-15,368)

Lucas, J. F. Personal communication, November 1973.

Lucas, J. F. The teaching of heuristic problem solving strategies in elementary calculus. *Journal for Research in Mathematics Education*, 1974, *5*, 36-46.

Luchins, A. S., & Luchins, E. H. New experimental attempts at preventing mechanization in problem solving. *Journal of General Psychology*, 1950, *42*, 279-297.

Maccoby, E. E. Sex differences in intellectual functioning. In E. E. Maccoby (Ed.), *The development of sex differences*. Stanford, CA: Stanford University, 1966.

Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, CA: Stanford University, 1974.

Maier, N. R. *Problem solving and creativity*. Belmont, CA: Brooks/Cole Pub. Co., 1931.

Mandler, G., & Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, *47*, 166-173.

Mastantuono, A. K. An examination of four arithmetic attitude scales (Doctoral dissertation, Case Western Reserve University, 1970). *Dissertation Abstracts International*, 1971, *32*, 248A.

Maxwell, A. A. An exploratory study of secondary school geometry students: Problem solving related to convergent-divergent productivity (Doctoral dissertation, University of Tennessee, 1974). *Dissertation Abstracts International*, 1975, *35*, 4987A. (University Microfilms No. 75-3,626)

May, D. C. An investigation of the relationship between selected personality characteristics of eighth-grade students and their achievement in mathematics (Doctoral dissertation, University of Florida, 1971). *Dissertation Abstracts International*, 1972, *33*, 555A.

May, R. *Man's search for himself*. New York: Signet Books Paperback Edition, New American Library, Inc., 1967. (Originally published, 1953.)

McClintock, C. E. *The hand calculator as an algebra I problem solving tool*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, San Diego, April 1978.

McKeachie, W. J. Student-centered versus instructor-centered instruction. *Journal of Educational Psychology*, 1954, *45*(3), 143-150.

McKeen, R. L. *A model for curriculum construction through observations of students solving problems in small instructional groups*. Unpublished doctoral dissertation, University of Maryland, 1970.

McKeen, R., & Davidson, N. An alternative to individual instruction in mathematics. *American Mathematical Monthly*, December 1975, 1006-1009.

McLeod, G. K., & Kilpatrick, J. Patterns of mathematics achievement in grades 7 and 8: Y-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 12). Stanford, CA: School Mathematics Study Group, Stanford University, 1969.

McLeod, G. K., & Kilpatrick, J. Patterns of mathematics achievement in grade 10: Y-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 14). Stanford, CA: School Mathematics Study Group, Stanford University, 1971.

McTaggart, H. P., & Jacklin, C. M. *The psychology of sex differences*. Stanford, CA: Stanford University, 1974.

Menchinskaya, N. A. [Intellectual activity in solving arithmetic problems.] In J. Kilpatrick & I. Wirszup (Eds.) & D. A. Henderson (trans.), *Soviet studies in the psychology of learning and teaching psychology* (Vol. III). Stanford, CA: School Mathematics Study Group and Survey of Recent East European Mathematical Literature, University of Chicago, 1969. (Originally published, 1946.)

Meyer, R. A. *A study of the relationship of mathematical problem solving performance and intellectual abilities of fourth-grade boys and girls* (Working Paper 160). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1976.

Michael, W. B., Guilford, J. P., Fruchter, B., & Zimmerman, W. S. The description of spatial-visualization abilities. *Educational and Psychological Measurement*, 1957, *17*, 185-199.

Mills, T. M. Power relations in three-person groups. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory* (2nd ed.). New York: Harper & Row, 1960.

Milton, G. A. The effects of sex-role identification upon problem solving skill. *Journal of Abnormal and Social Psychology*, 1957, *55*, 219-244.

Milton, G. A. *Five studies of the relation between sex role identification and achievement in problem solving* (Technical Report 3). New Haven, CT: Department of Psychology and Department of Industrial Administration, Yale University, 1958.

Mitchelmore, M. C. *Cross-cultural research on concepts of space and geometry*. Paper presented at the Research Workshop on Space and Geometry, Georgia, May 1975.

Moise, E. E. Activity and motivation in mathematics. *American Mathematical Monthly*, 1965, *72*(4), 407-412.

Montagu, A. *On being human*. New York: Hawthorn, 1966.

Montgomery, M. E., & Whitaker, D. R. *Report of the coordinators' training for large scale field testing of Developing Mathematical Processes* (Technical Report 296). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1975.

Moore, B. D. The relationship of fifth-grade students' self-concepts and attitudes toward mathematics to academic achievement in arithmetical computation, concepts, and application (Doctoral dissertation, North Texas State University, 1971). *Dissertation Abstracts International*, 1972, *32*, 4426A.

Moses, B. *The role of spatial characteristics in the problem solving process*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, San Diego, April 1978.

National Advisory Committee on Mathematical Education. *Overview and analysis of school mathematics K-12*. Washington, D.C.: Author, 1975.

National Council of Supervisors of Mathematics. *Position paper on basic mathematical skills*. Prepared for National Institute of Education, U.S. Department of Health, Education, and Welfare, July 1976.

Naylor, F. D., & Gaudry, E. The relationship of adjustment, anxiety, and intelligence to mathematics performance. *Journal of Educational Research*, 1973, *66*, 413-417.

Nelson, L. D., & Kilpatrick, J. Problem solving. In J. Payne (ed.), *Mathematics learning in early childhood, thirty-seventh yearbook, National Council of Teachers of Mathematics*. Reston, VA: National Council of Teachers of Mathematics, Inc., 1975.

Neufeld, K. A. Differences in personality characteristics between groups having high and low mathematical achievement gain under individualized instruction (Doctoral dissertation, University of Wisconsin, 1967). *Dissertation Abstracts International*, 1968, *28*, 4540A.

Newell, A., Shaw, J. C., & Simon, H. A. Elements of a theory of human problem solving. *Psychological Review*, 1958, *65*, 151-166.

Norman, P. B. *Relationships between problem-solving ability, computational skill, intelligence, and amount of training in mathematics*. Unpublished doctoral dissertation, Columbia University, 1950.

Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1976.

Ontario Institute for Studies in Education. *K-13 mathematics, some non-geometric aspects, part II: Computing, logic, and problem solving.* Toronto: Author, 1971.

Ortiz-Franco, L. A selected study on mathematical word solving processes (Doctoral dissertation, Stanford University, 1977). *Dissertation Abstracts International,* 1978, *38,* 5312A. (University Microfilms No. 78-0,211)

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning.* Urbana: University of Illinois, 1957.

Otis, A. S., & Lennon, R. T. *Otis-Lennon mental ability test.* New York: Harcourt, Brace, Jovanovich, 1970.

Parsley, K. M., Powell, M., & O'Connor, H. A. Further investigation of sex differences in achievement of under-, average- and over-achieving students within five IQ groups in grades four through eight. *Journal of Educational Research,* 1964, *57,* 268-270.

Parsley, K. M., Powell, M., O'Connor, H. A., & Deutsch, M. Are there really sex differences in achievement? *Journal of Educational Research,* 1963, *56,* 210-212.

Pennington, B. A. Behavioral and conceptual strategies as decision models for solving problems (Doctoral dissertation, University of California, Los Angeles, 1970). *Dissertation Abstracts International,* 1970, *31,* 1630-1631A. (University Microfilms No. 70-19,879)

Pereira-Mendoza, L. The effect of teaching heuristics on the ability of grade ten students to solve novel mathematical problems (Doctoral dissertation, University of British Columbia, 1975). *Dissertation Abstracts International,* 1976, *36,* 8014A. (a)

Pereira-Mendoza, L. *The effect of instruction in heuristics on the ability to solve open ended mathematical problems.* Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Atlanta, April 1976. (b)

Phillips, R. B. Teacher attitude as related to student attitude and achievement in elementary school mathematics. *School Science and Mathematics,* 1973, *73,* 501-507.

Poffenberger, T. Research note on father-child relations and father viewed as negative figure. *Child Development,* 1959, *30,* 489-492.

Poffenberger, T., & Norton, D. A. Factors in the formation of attitudes toward mathematics. *Journal of Educational Research,* 1959, *52,* 171-176.

Poincaré, H. *The foundations of science*. New York and Lancaster: The Science Press, 1929.

Poincaré, H. [Intuition and logic in mathematics] (G. B. Halstead, trans.). *Mathematics Teacher*, 1969, *62*(3), 205-212. (Originally published in English in 1929.) (a)

Poincaré, H. Mathematical definitions and teaching. *Mathematics Teacher*, 1969, *62*(4), 295-304. (b)

Polya, G. *How to solve it*. Princeton, NJ: Princeton University, 1945.

Polya, G. *How to solve it*. Princeton, NJ: Princeton University, 1948.

Polya, G. Induction and analogy in mathematics. In *Mathematics and plausible reasoning* (Vol. 1). Princeton, NJ: Princeton University, 1954. (a)

Polya, G. Patterns of plausible inference. In *Mathematics and plausible reasoning* (Vol. 2). Princeton, NJ: Princeton University, 1954. (b)

Polya, G. *How to solve it* (2nd ed.). Garden City, NY: Doubleday Anchor Books, 1957.

Polya, G. *Lecture notes for problem solving seminar*. Stanford, CA: Stanford University, 1960.

Polya, G. *Mathematical discovery: On understanding, learning, and teaching problem solving* (Vol. I). New York: John Wiley & Sons, 1962.

Polya, G. On learning, teaching, and learning teaching. *American Mathematical Monthly*, 1963, *70*, 605-619.

Polya, G. *Mathematical discovery: On understanding, learning, and teaching problem solving* (Vol. II). New York: John Wiley & Sons, 1965.

Poppendieck, M. *The arithmetic algorithms*. Unpublished master's thesis, University of Maryland, 1971.

Post, T. R., & Brennan, M. L. An experimental study of the effectiveness of a formal versus an informal presentation of a general heuristic process on problem solving in tenth-grade geometry. *Journal for Research in Mathematics Education*, 1976, *7*, 59-64.

Reys, R. E., & Delon, F. G. Attitudes of prospective elementary school teachers toward arithmetic. *Arithmetic Teacher*, 1968, *15*(4), 363-366.

Richards, C. M. Third thoughts on discovery. *Educational Review* (British), 1973, *25*, 143-150.

Robertson, H. C. The effects of the discovery and expository approach of presenting and teaching selected mathematical principles and relationships to fourth-grade pupils (Doctoral dissertation, University of Pitts-

burgh, 1970). *Dissertation Abstracts International*, 1971, *31*, 5278A-5279A. (University Microfilms No. 71-8, 785)

Robinson, M. L. An investigation of problem solving behavior and cognitive and affective characteristics of good and poor problem solvers in sixth-grade mathematics (Doctoral dissertation, State University of New York at Buffalo, 1973). *Dissertation Abstracts International*, 1973, *33*, 5620A. (University Microfilms No. 73-9, 745)

Romberg, T. A. Current research in mathematics education. *Review of Educational Research*, 1969, *39*, 473-491.

Romberg, T. A. *Specifications for terminal accountability tests for DMP* (Project Paper 74-1). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1974.

Romberg, T. A. *Romberg mathematics computation test*. Madison: Wisconsin Research and Development Center for Cognitive Learning, 1975.

Romberg, T. A. Developing mathematical processes. In H. J. Klausmeier, R. E. Rossmiller, & M. H. Saily, *Individually guided education*. New York: Academic Press, 1976.

Romberg, T. A., & DeVault, M. V. Mathematics curriculum: Needed research. *Journal of Research and Development in Education*, 1967, *1*(1), 95-112.

Romberg, T. A., & Harvey, J. G. *Developing mathematical processes: Background and projections* (Working Paper 14). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1969.

Romberg, T. A., & Harvey, J. G., Moser, J. M., & Montgomery, M. E. *Developing mathematical processes*. Chicago: Rand McNally & Co., 1974, 1975, 1976.

Romberg, T. A., Harvey, J. G., Moser, J. M., Montgomery, M. E., & Dana, M. E. *Developing mathematical processes*. Chicago: Rand McNally & Co., 1974, 1975, 1976.

Romberg, T. A., & Wilson, J. W. The development of tests. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 7). Stanford, CA: School Mathematics Study Group, Stanford University, 1969.

Rubinstein, M. *Patterns of problem solving*. Englewood Cliffs, NJ: Prentice-Hall, 1974.

Sarnoff, I. *Society with tears*. New York: Citadel Press, 1966.

Scheffe, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, *40*, 87-104.

Schwartz, B. L. A new sliding block puzzle. *Mathematics Teacher*, 1973, *66*, 277-280.

Schwieger, R. D. A component analysis of mathematical problem solving (Doctoral dissertation, Purdue University, 1974). *Dissertation Abstracts International*, 1974, *35*, 3308-3309A. (University Microfilms No. 74-26,777)

Science Research Associates, Inc. *Primary mental abilities tests*. Chicago: Author, 1962.

Science Research Associates, Inc. *Modern math understanding test form C, multilevel edition*. Chicago: Author, 1966.

Scott, J. A., & Frayer, D. A. *Learning by discovery: A review of the research methodology* (Working Paper 64). Madison: Wisconsin Research and Development Center for Cognitive Learning, December 1970. (ERIC Document Reproduction Service No. ED 053 793)

Scott, W. A. Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed.). Reading, MA: Addison-Wesley, 1968.

Seeley, R. T. *Calculus of one variable* (2nd ed.). Glenview, IL: Scott Foresman & Co., 1972.

Seidl, N. W. *An application of a small group instructional method to identify some adaptations of curriculum and instruction for homogeneous groups*. Unpublished doctoral dissertation, University of Maryland, 1971.

Sells, L. W. *High school math as a critical filter in the job market*. March 31, 1973, 6 p. (ERIC Document Reproduction Service No. ED 080 351)

Shapiro, E. W. Attitudes toward arithmetic among public school children in the intermediate grades (Doctoral dissertation, University of Denver, 1961). *Dissertation Abstracts International*, 1962, *22*, 3927-3928.

Sheehan, T. J. Patterns of sex differences in learning mathematical problem solving. *Journal of Experimental Education*, 1968, *36*, 84-87.

Shepler, J. L. *A study of parts of the development of a unit in probability and statistics for the elementary school* (Technical Report 105). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1969.

Sheridan Psychological Services, Inc. *Figure matrix*. Beverly Hills, CA: Author, 1969.

Sherman, J. A. Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review*, 1967, *4*, 290-299.

Sherman, J. A. Field articulation, sex, spatial visualization, dependency, practice, laterality of the brain and birth order. *Perceptual and Motor Skills*, 1974, *38*, 1223-1235.

Sherman, J., & Fennema, E. The study of mathematics by high school girls and boys: Related variables. *American Educational Research Journal*, 1977, *14*, 159-168.

Sherrill, J. M. The effects of different presentations of mathematical word problems upon achievement of tenth-grade students. *School Science and Mathematics*, 1973, *73*, 277-282.

Shriner, D. F. *An attempt to identify curricular and instructional variations for students of different ability levels: An exploratory application of small group curriculum development methods.* Unpublished doctoral dissertation, University of Maryland, College Park, 1970.

Shulman, L. S. Psychology and mathematics education. In E. G. Begle (Ed.), *Mathematics education*. Chicago: National Society for the Study of Education, 1970.

Silver, E. A. *An examination of student perceptions of relatedness among mathematical word problems*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, San Diego, April 1978.

Slater, P. E. Contrasting correlates of group size. *Sociometry*, 1958, *21*(2), 129-139.

Smith, I. M. *Spatial ability*. London: University of London Press Ltd., 1964.

Smith, J. P. The effect of general versus specific heuristics in mathematical problem solving tasks (Doctoral dissertation, Columbia University, 1973). *Dissertation Abstracts International*, 1973, *34*, 2400A. (University Microfilms No. 73-26,637)

Stafford, R. E. Sex-differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and Motor Skills*, 1961, *13*, 428.

Stanley, J. C., Keating, D. P., & Fox, L. H. *Mathematical talent: Discovery, description and development*. Baltimore: Johns Hopkins University Press, 1974.

Stein, S., & Crabill, C. *Elementary algebra, a guided inquiry*. Boston: Houghton Mifflin Co., 1972.

Stright, V. M. A study of the attitudes toward arithmetic of students and teachers in the third, fourth, and sixth grades. *Arithmetic Teacher*, 1960, *7*, 280-286.

Swafford, J. O. A study of the relationship between personality and achievement in mathematics (Doctoral dissertation, University of Georgia, 1969). *Dissertation Abstracts International*, 1970, *30*, 5353A.

Sweeney, E. J. *Sex differences in problem solving* (Technical Report 1). Stanford, CA: Stanford University Department of Psychology, 1953.

Symonds, P. M. What education has to learn from psychology. *Teacher's College Record*, 1958, *60*(1), 9-22.

Tanner, R. T. Discovery as an object of research. *School Science and Mathematics*, 1969, *69*, 647-655.

Tate, M. W., & Stanier, B. Errors in judgment of good and poor problem solvers. *Journal of Experimental Education*, 1964, *32*, 371-376.

Thompson, E. N. Readability and accessory remarks: Factors in problem solving in arithmetic (Doctoral dissertation, Stanford University, 1967). *Dissertation Abstracts International*, 1968, *28*, 2464A-2465A.

Thoyre, H. *A pilot study on the use of small group discussion in a mathematics course for preservice elementary teachers.* Unpublished doctoral dissertation, University of Wisconsin, 1970.

Thurstone, L. L. Attitudes can be measured. *American Journal of Sociology*, 1928, *33*, 529-554.

Thurstone, L. L., & Jiffrey, T. E. *Gottschaldt concealed figures.* Chicago: Industrial Relations Center, University of Chicago, 1956.

Thurstone, T. G. *Manual for the SRA primary mental abilities tests — ages 11 to 17.* Cleveland: Science Research Associates, 1958.

Tiegs, E. W., & Clark, W. W. *California achievement tests: Complete battery.* Monterey, CA: CTB/McGraw-Hill, 1970.

Todd, R. M. A mathematics course for elementary teachers: Does it improve understanding and attitudes? *Arithmetic Teacher*, 1966, *13*, 198-202.

Torgerson, W. J. *Theory and methods of scaling.* New York: John Wiley & Sons, 1958.

Torrance, E. P. *Characteristics of mathematics teachers that affect students' learning* (Report No. CRP-1020). Washington, D.C.: U.S. Office of Education, 1966.

Treacy, J. P. The relationship of reading skills to the ability to solve arithmetic problems. *Journal of Educational Research*, 1944, *38*, 86-95.

Tubb, G. W. Heuristics questioning and problem solving strategies in mathematics graduate teaching assitants and their students (Doctoral disserta-

tion, Texas A & M University, 1974). *Dissertation Abstracts International*, 1975, *36*, 235-236A. (University Microfilms No. 75-15,077)

Turner, V. D., Alders, C. D., Hatfield, F., Croy, H., & Sigrist, C. A study of ways of handling large classes in freshman mathematics. *American Mathematical Monthly*, 1966, *73*(7), 768-770.

Tyler, L. E. *The psychology of human differences*. New York: Appleton-Century-Crofts, 1965.

University of Maryland Mathematics Project (UMMAP). *Unifying concepts and processes in elementary mathematics*. Boston: Allyn & Bacon, Inc., 1978.

Vanderlene, L. F. Does the study of quantitative vocabulary improve problem solving? *Elementary School Journal*, 1964, *65*, 143-152.

Van de Walle, J. A. Attitudes and perceptions of elementary mathematics possessed by third and sixth grade teachers as related to student attitude and achievement in mathematics (Doctoral dissertation, Ohio State University, 1972). *Dissertation Abstracts International*, 1973, *33*, 4254A-4255A.

Very, P. S. Differential factor structures in mathematical ability. *Genetic Psychology Monographs*, 1967, *75*, 169-207.

Vos, K. E. The effects of three instructional strategies on problem solving behaviors in secondary school mathematics. *Journal for Research in Mathematics Education*, 1976, *7*, 264-275.

Vos, K. E. *The effects of three key organizers on mathematical problem solving success with sixth-, seventh-, and eighth-grade learners*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, San Diego, April 1978.

Wearne, D. C. *Development of a test of mathematical problem solving which yields a comprehension, application, and problem solving score* (Technical Report 407). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1976.

Webb, L. F., & Sherrill, J. M. The effects of differing presentations of mathematical word problems upon the achievement of preservice elementary teachers. *School Science and Mathematics*, 1974, *74*, 559-565.

Webb, N. L. An exploration of mathematical problem solving processes (Doctoral dissertation, Stanford University, 1975). *Dissertation Abstracts International*, 1975, *36*, 2689A. (University Microfilms No. 75-25,625) (a)

Webb, N. L. *An exploration of mathematical problem solving processes.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975. (b)

Weissglass, J. Small groups: An alternative to the lecture method. *The Two-Year College Mathematics Journal,* 1976, *VII,* 15-20.

Weissglass, J. Mathematics for elementary teaching: A small-group laboratory approach. *American Mathematical Monthly,* May 1977, 377-382.

Weissglass, J. *Exploring elementary mathematics: A small group approach for teaching.* San Francisco: W. H. Freeman, 1979.

Werdelin, I. *The mathematical ability.* Lund, Sweden: C. W. K. Gleerup, 1958.

Werdelin, I. *Geometrical ability and the space factor in boys and girls.* Lund, Sweden: C. W. K. Gleerup, 1961.

Werdelin, I. A synthesis of two factor analyses of problem solving in mathematics. *Didokometary,* 1966, *8,* 1-13 and Tables 14-23.

Wess, R. G. An analysis of the relationship of teachers' attitudes as compared to pupils' attitudes and achievement in mathematics (Doctoral dissertation, University of South Dakota, 1969). *Dissertation Abstracts International,* 1970, *30,* 3844A-3845A.

Whitaker, D. R. *A study of the relationships between selected noncognitive factors and the problem solving performance of fourth-grade children* (Technical Report 396). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1976.

White, M. J. A. A study of the change of achievement and attitude toward arithmetic by prospective elementary school teachers under the conditions of television (Doctoral dissertation, Wayne State University, 1963). *Dissertation Abstracts International,* 1965, *25,* 2302-2303.

White, R., & Lippitt, R. Leader behavior and member reaction in three "social climates." In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory.* New York: Harper & Row, 1960.

Whyburn, L. S. Student-oriented teaching — the Moore method. *American Mathematical Monthly,* 1970, *77*(4), 351-359.

Wickelgren, W. *How to solve problems: Elements of a theory of problems and problem solving.* San Francisco: W. H. Freeman, 1974.

Wickes, H. E. Pre-service mathematics preparation of elementary teachers: The relative effectiveness of two programs in determining attitudes toward and achievement in mathematics (Doctoral dissertation, Colo-

rado State College, 1967). *Dissertation Abstracts International*, 1968, *28*, 2591A.

Willoughby, S. S. Mathematics. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.). New York: Macmillan, 1969.

Wilson, J. W. Patterns of mathematics achievement in grade 10: Z-population. In J. W. Wilson & E. G. Begle (Eds.), *NLSMA Reports* (No. 16). Stanford, CA: School Mathematics Study Group, Stanford University, 1972. (a)

Wilson, J. W. Patterns of achievement in grade 11: Z-population. In J. W. Wilson, L. S. Cahen, & E. G. Begle (Eds.), *NLSMA Reports* (No. 17). Stanford, CA: School Mathematics Study Group, Stanford University, 1972. (b)

Wilson, J. W. Analysis of reasoning processes. In J. Kilpatrick, I. Wirszup, E. G. Begle, & J. W. Wilson (Eds. and trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. XIII). Chicago: University of Chicago, 1975.

Wilson, J. W., & Begle, E. G. (Eds.). Correlates of mathematics achievement: Summary. *NLSMA Reports* (No. 26). Stanford, CA: School Mathematics Study Group, Stanford University, 1972. (a)

Wilson, J. W., & Begle, E. G. (Eds.). Intercorrelations of mathematical and psychological variables. *NLSMA Reports* (No. 33). Stanford, CA: School Mathematics Study Group, Stanford University, 1972. (b)

Wilson, J. W., Cahen, L. S., & Begle, E. G. (Eds.). Z-population test batteries. *NLSMA Reports* (No. 3). Stanford, CA: School Mathematics Study Group, Stanford University, 1968. (a)

Wilson, J. W., Cahen, L. S., & Begle, E. G. (Eds.). Description and statistical properties of Z-population scales. *NLSMA Reports* (No. 6). Stanford, CA: School Mathematics Study Group, Stanford University, 1968. (b)

Wilson, J. W., Cahen, L. S., & Begle, E. G. (Eds.). Non-test data. *NLSMA Reports* (No. 9). Stanford, CA: School Mathematics Study Group, Stanford University, 1968. (c)

Wilson, J. W., Cahen, L. S., & Begle, E. G. (Eds.). Y-population test batteries. *NLSMA Reports* (No. 2A). Stanford, CA: School Mathematics Study Group, Stanford University, 1968. (d)

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Witkin, H. A. Individual differences in ease of perception of embedded figures. *Journal of Personality*, 1950, *19*, 1-15.

Witkin, H. A., Dyke, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological differentiation*. New York: John Wiley & Sons, 1962.

Wittrock, M. C. The learning by discovery hypothesis. In L. S. Shulman & E. R. Keislar (Eds.), *Learning by discovery*. Chicago: Rand McNally & Co., 1966.

Worthen, B. R. *Discovery vs. expository classroom instruction: An investigation of teaching mathematics in the elementary school*. Unpublished master's thesis, University of Utah, 1965.

Worthen, B. R. A study of discovery and expository presentations: Implications for teaching. *Journal of Teacher Education*, 1968, *19*, 223-242.

Yakimanskaya, I. S. [Individual differences in solving geometry problems on proof.] In J. Kilpatrick & I. Wirszup (Eds.) & J. W. Teller (trans.), *Soviet studies in the psychology of learning and teaching mathematics* (Vol. IV). Stanford, CA: School Mathematics Study Group, Stanford University, and Survey of Recent East European Mathematical Literature, University of Chicago, 1970. (Originally published, 1959.)

Zalewski, D. L. *An exploratory study to compare two performance measures: An interview-coding scheme of mathematical problem solving and a written test* (Technical Report 306). Madison: Wisconsin Research and Development Center for Cognitive Learning, 1974.